



UNIVERSITÀ DEGLI STUDI DI UDINE

Dottorato di Ricerca in Scienze e Biotecnologie Agrarie

Ciclo XXIX

Coordinatore: prof. Giuseppe Firrao

TESI DI DOTTORATO DI RICERCA

**Identification of structural variation in *Zea mays*:
use of paired-end mapping and development of a
novel algorithm based on split reads**

DOTTORANDO
Ettore Zapparoli

SUPERVISORE
Prof. Michele Morgante

CO-SUPERVISORE
Dott. Fabio Marroni

ANNO ACCADEMICO 2015/2016

Contents

SUMMARY	I
1 INTRODUCTION	1
1.1 <i>Maize and its genome</i>	1
1.2 <i>Structural variants and maize pan-genome</i>	5
1.3 <i>Methods for the detection of structural variation analysis</i>	10
2 OBJECTIVES	13
3 MATERIALS AND METHODS	14
3.1 <i>Library preparation and sequencing</i>	14
3.2 <i>Alignment and SNP calling</i>	15
3.3 <i>Identification of deletions</i>	16
3.4 <i>Transposable elements annotation</i>	18
3.5 <i>Dating of Long Terminal Repeat insertion events</i>	19
3.6 <i>Walle algorithm</i>	20
3.7 <i>Detection of insertions on simulated data</i>	28
3.8 <i>Detection of insertions in maize</i>	30
3.9 <i>Analysis of genes affected by SV</i>	32
3.10 <i>Validation of SVs on de novo assembly</i>	33
4 RESULTS AND DISCUSSION	35
4.1 <i>Sequencing</i>	35
4.2 <i>Software development for the identification of insertions</i>	37
4.3 <i>Identification of deletions in Zea mays</i>	41
4.4 <i>Identification of insertions in Zea mays</i>	48
4.5 <i>Classification of SV</i>	55
4.6 <i>Nested elements analysis</i>	65
4.7 <i>SV validation based on de novo assembly</i>	70
5 CONCLUSIONS	72
REFERENCES	76
ACKNOWLEDGMENTS	84

SUMMARY

The present work is part of the ERC-funded project NOVABREED, which has as objective the characterization of the pan-genome of *Vitis vinifera* and *Zea mays*, through the application and the development of *in silico* methods for the analysis of Next Generation Sequencing (NGS) data.

The concept of pan-genome arises from the observation that some DNA sequences are not shared by all subjects of a species, and that a single genome is not enough to describe the species. The DNA segments shared from all subjects of a species constitute the core genome, while those not present in all subjects compose the dispensable genome. Here, we focused on the genome of *Zea mays*, a complex and highly repeated genome, whose size is approximately 2.5 Gb (Schnable et al., 2009).

Structural variants are an important source of genetic variation in plants, mostly due to large (>1000 bp) insertions and deletions of transposable elements (TEs) and are an important component of the dispensable genome. Maize dispensable fraction of the genome was characterized through the analysis of structural variants (SVs) in 7 inbred lines selected from the parental lines of the MAGIC maize population.

As part of the project, a new algorithm (Walle) for the detection of insertions relying on split-read mapping (SR) has been developed, and its performance has been compared with existing tools. Results showed that Walle performed better than existing tools.

Deletions were detected using publicly available tools, while insertions were detected using tools previously detected in our lab and the tool developed in the present project.

A total of 48,904 deletions and 75,370 insertions were identified, accounting respectively for 0.56 Gb of sequences present in the B73 reference genome and absent in at least one other line, and an estimated 0.81 Gb of sequences present in at least one other line while absent in B73.

The composition of dispensable genome was investigated, confirming that a large fraction of extant variation in maize is due to LTR retrotransposons insertions and that most of them occurred in a relatively recent time.

Although most SVs are located in intergenic regions, some of them are located in genes and may disrupt exons, leading to evolutionary consequences. We therefore assessed the function of genes affected by deletions and insertions.

Nested elements were investigated in greater detail, and we confirmed that LTR retrotransposons form nesting structures more often than expected by chance alone, as previously reported (Jiang and Wessler, 2001). Moreover, nesting patterns were investigated, finding that most of nesting events occurs within a few families of LTR retrotransposons.

The main results of the present work are a) a software tool for the accurate identification of insertions in the genome, which has been shown to outperform existing tools, has been used for the identification of insertions in *Zea mays* and can be used on the genome of any species, and b) the characterization of the dispensable genome of *Zea mays*, which resulted in important information on the patterns of the movement of transposable elements, on their nesting patterns, and on the function of genes affected by the movement of TEs.

1 INTRODUCTION

1.1 Maize and its genome

Zea mays (or maize) is an annual monocotyledon outcrossing species, propagated by seed, and characterized by highly homozygous individuals (commercial inbreds) and highly heterozygous ones (commercial F1 hybrids).

The development of human agriculture is mostly responsible of the domestication of maize from the selective breeding of mexican grass teosinte, from ~10,000 to ~4,000 years ago (Doebley et al., 2006).

Today, maize is an important crop for the production of food, feed and ethanol for biofuel, yielding 1 billion tons worldwide (2014 data from <http://faostat.fao.org/>), and in the USA alone 360 million tons from ~90.6 million acres with a value of \$52 B (2014 data from <http://ncga.com/>).

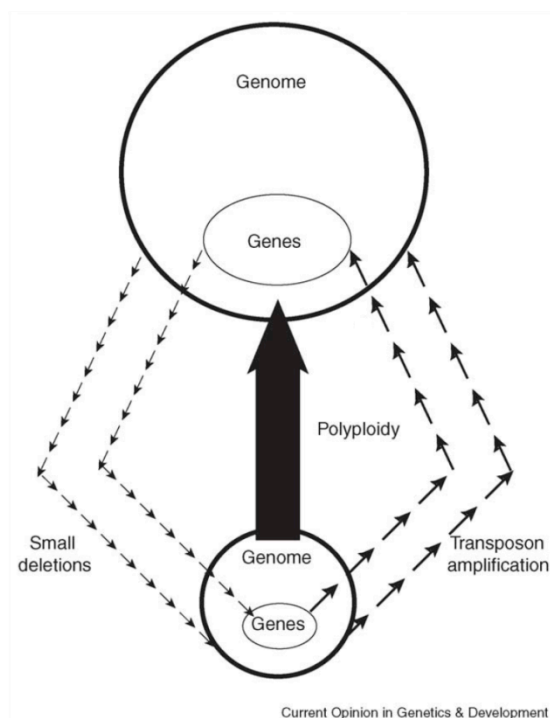


Figure 1. The process leading to quantitative increasing of DNA content in plant genomes (Bennetzen, *Curr Opin Genet Dev.* 2005; reproduced with permission).

Even though the maize genome has a diploid set of chromosomes, there are evidences of its allotetraploid origin, as a result of whole genome duplication event between 5 and 12 million of years ago due to the hybridization of two maize progenitors (Bruggmann et al., 2006), after their divergence from the ancestral sorghum ~12 million of years ago (Swigoňová et al., 2004); this event is peculiar for the maize genome, which already underwent the ancient whole-genome duplication common to all grass species (Paterson et

al., 2004), and preceded the LTR retrotransposons proliferation, which is estimated to have expanded the maize genome size from 1.2 gigabases to 2.4 gigabases in the last 3 million of years (SanMiguel et al., 1998; Bennetzen, 2005; Piegu et al., 2006).

Hence, maize has a quite large genome (2.4 gigabases) with 10 chromosomes and ~39,000 genes. The first BAC-by-BAC assembly of B73 reference genome (RefGen_v1) was released in 2009 (Schnable et al., 2009).

Two other versions were released in 2010 (RefGen_v2) and in 2013 (RefGen_v3), to fill in gaps in previous assemblies via WGS and re-orientate contigs via Sorghum-guided synteny.

An early release of a new assembly (RefGen_v4) is available to date, which relies on PacBio Single Molecule Real Time (SMRT) sequencing at 60X coverage. Although very promising, this is an early release of data and curators declares that “The underlying assembly and final set of annotations may change until the data has been accepted by GenBank”.

Almost 85% of the maize genome is composed by transposable elements (TE), with a >75% component of LTR class I RNA retrotransposons Gypsy and Copia (Baucom et al., 2009), and the remainder distributed among other class I retrotransposons (LINE), and class II DNA elements such as CACTA (Spm/En), hAT (Ac), PIF/Harbinger, Mutator (Mu), MITE (Tourist) and Helitrons (Tenaillon, 2011).

A unified classification system of such transposable elements was proposed by Wicker et al. in 2007 (Wicker et al., 2007), with a 3-letters code prepended to the TE family name.

Class I “copy and paste” RNA retrotransposons transpose using a reverse transcription mechanism and are present in maize mainly as:

- LTR class elements (fig. 2b), which have a long-terminal repeat sequence flanking the internal region, where are encoded genes for the transposition activity (Havecker et al., 2004)
- L1 and RTE superfamilies of long interspersed nuclear element (LINE) class, non-LTR elements (fig. 2c) which typically present a poly(A) tail at 3', while during the reverse transcription a 5' truncated copy can be generated (Eickbush and Malik, 2002)

Class II “cut and paste” DNA transposons are present mainly as:

- Terminal inverted repeat (TIR) class elements (fig. 2a), in particular superfamilies of hAT, Mutator, Harbinger and CACTA elements, having a terminal inverted repeats (TIRs) region of different size (Gierl and Frey, 1991)
- Helitron elements, which transpose through a rolling-circle mechanism via a single-stranded DNA intermediate (fig. 3), thanks to a Helicase complex (Rep Helicase) encoded by all autonomous Helitron elements (Morgante et al., 2005; Li and Dooner, 2009)

Due to high number of TEs in the maize genome, the formation of nested TEs (i.e. the insertion of a TE into another TE) is rather common. In *Zea mays* such events were first observed for the more abundant retrotransposons belonging to the LTR superfamilies *Gypsy* and *Copia* (SanMiguel et al., 1996), then studied genome-wide including other superfamilies such as Helitrons, LINE and TIR (Kronmiller and Wise, 2009; Gao et al., 2012).

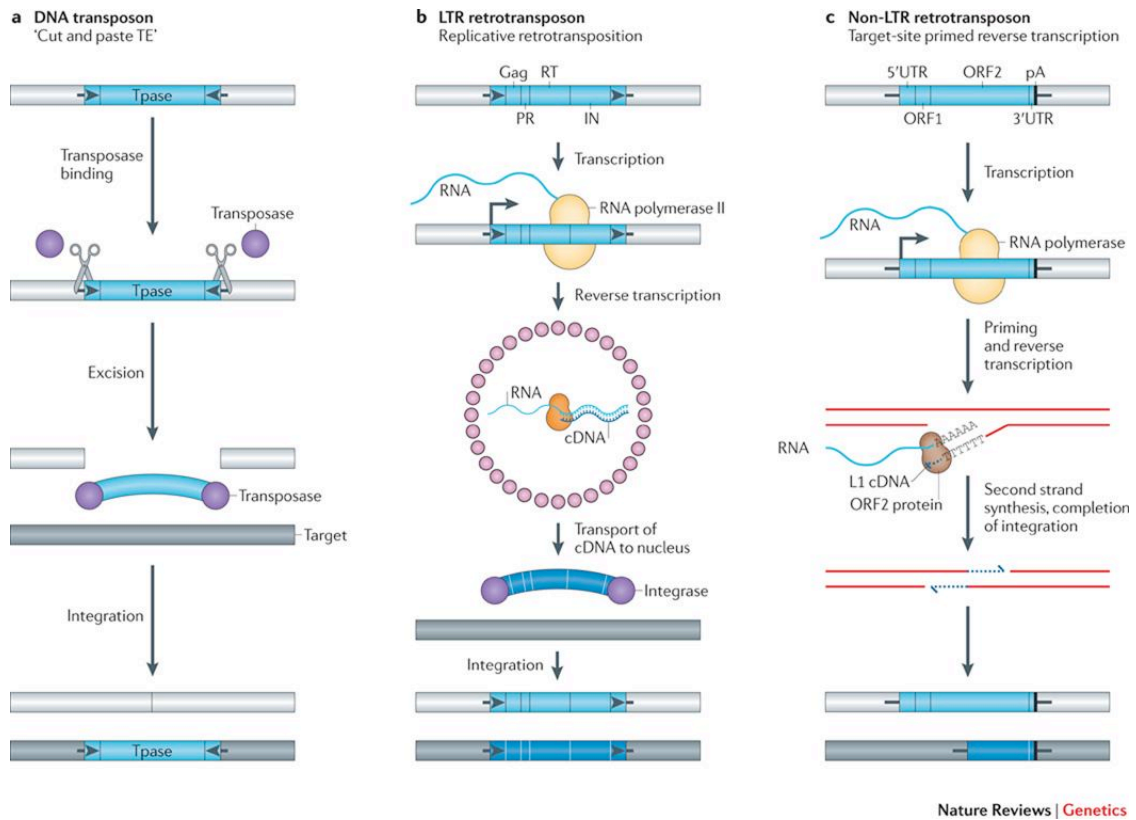


Figure 2. Model for the transposition of Class I retrotransposon elements (b,c) and TIR class II DNA elements (a) (Levin and Moran, *Nat Rev Genetics* 2011; reproduced with permission).

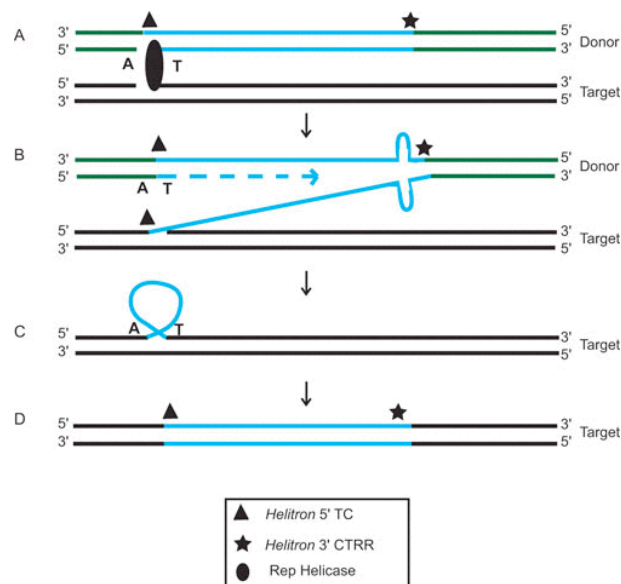


Figure 3. Model for the transposition of Helitrons (Thomas and Pritham, *Microbiol Spectrum* 2014; reproduced with permission).

Transposition of TEs was discovered in maize by Nobel prize laureate Barbara McClintock, who also showed that TEs can heavily influence gene expression (McClintock, 1951).

Transposable elements have been annotated and recorded in databases of TEs that may be easily accessed by researchers. The Maize transposable element database (Wessler et al., 2009, <http://maizetdb.org>) is the most comprehensive repository of TEs identified in maize and represents an extremely valuable resource for the annotation of TEs.

1.2 Structural variants and maize pan-genome

The large amount of TEs in the maize genome directly affects intraspecies diversity, as TE transpositions can produce structural variants (SVs), i.e. modifications involving DNA sequences longer than 1 Kb (Feuk et al., 2006). Other mechanisms that can produce SVs are non-allelic homologous recombination (NAHR) between two regions with high sequence similarity, generated during ancient duplication events (Eichler and Sankoff, 2003). In contrast to genome expansion mediated by TE transposition, NAHR is more involved in genome fractionation, i.e. the loss of duplicate genes after whole genome duplication (Devos et al., 2002; Sankoff et al., 2012).

DNA sequencing, either through Sanger sequencing or through Next generation sequencing (NGS) technology, has helped in investigating the role of SV in contributing to genetic diversity in maize, and it was shown that a wide range of TEs contribute to diversity as presence/absence variation (PAV) in several loci between B73 and Mo17 inbred lines (Brunner et al., 2005; Eichten et al., 2011). Brunner and colleagues showed that 50% of sequences are shared between the two lines, while the remaining 50% are mainly composed by TEs, suggesting that TEs are the major source of SV in maize. Another analysis of the maize *bz* locus - already known to be a variable locus (Fu and Dooner, 2002) - in eight different

inbred lines confirmed the strong diversity in maize, due to TE-induced SVs (Wang et al., 2006).

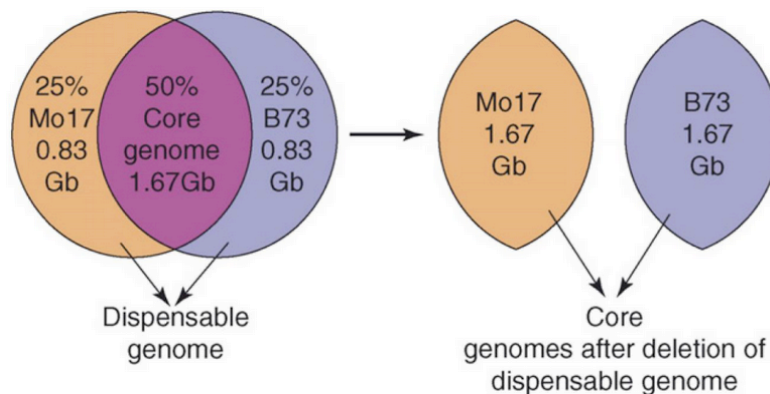


Figure 4. A pan-genome view of the maize genome as defined by comparison of sequenced genomic regions in the B73 and Mo17 inbred lines (Morgante et al., Curr Opin Plant Biol 2007; reproduced with permission).

Comparative genomic hybridization (CGH) array studies were also performed at a whole genome scale (Springer et al., 2009) between B73 and Mo17 inbred lines, showing a high number of copy number variants (CNV) and a higher number of PAVs present in B73 but not in Mo17, thus confirming previous observations. Another CGH array study showed that on a panel of 19 inbred maize lines, an extensive fraction of genes is affected by CNV and even more by PAV (Swanson-Wagner et al., 2010). The resequencing of six elite maize inbred lines (Lai et al., 2010) - included Mo17 - had identified several genes absent in Mo17 and in other varieties, for a total of 296 genes absent in at least one variety. Another work pointed out that on 27 diverse resequenced maize lines, B73 genome is estimated to capture only ~70% of the available low-copy sequences (Gore et al., 2009), while in a subsequent study 8681 representative novel transcripts absent in B73 were assembled (Hirsch et al., 2014). Moreover, it was shown that SVs might cause phenotypic variation (Chia et al., 2012), as they can be found in the proximity of genes (Tenaillon et al., 2010; Wang et al., 2013) or cause genic copy number variation (Maron et al., 2013).

Transposon insertions may be the cause of a large number of SVs causing phenotypic alteration in plants (Lisch et al., 2012). That was observed for specific cases, i.e. in maize, where it was studied how the insertion of a *Copia* LTR retrotransposon in a regulatory region 60 kb upstream of the coding sequence of *tb1* gene, results in an enhancer-like activity and consequent increase of gene expression (Studer et al., 2011). Conversely, recent subsequent retroelement insertions into the first exon of *b1* gene in maize *B-Bolivia* allele caused phenotypic effects, resulting in reduced and variegated expression of that gene (Selinger et al., 2001). In *V. vinifera* the insertion of the LTR retrotransposon *GRET1* may influence an important biosynthetic pathway, influencing the color of berries (Kobayashi et al., 2004). Similar mechanisms were also found in blood oranges (Butelli et al., 2012), while in peach seems to disrupt genes (Falchi et al., 2013).

The ability of helitron elements to capture gene fragments was observed in maize (Yang and Bennetzen, 2009; Du et al., 2009), and their ability to lead to exon shuffling was documented (Morgante et al., 2005). However, such elements are still difficult to detect genome-wide due to their atypical structure (Thomas and Pritham, 2015).

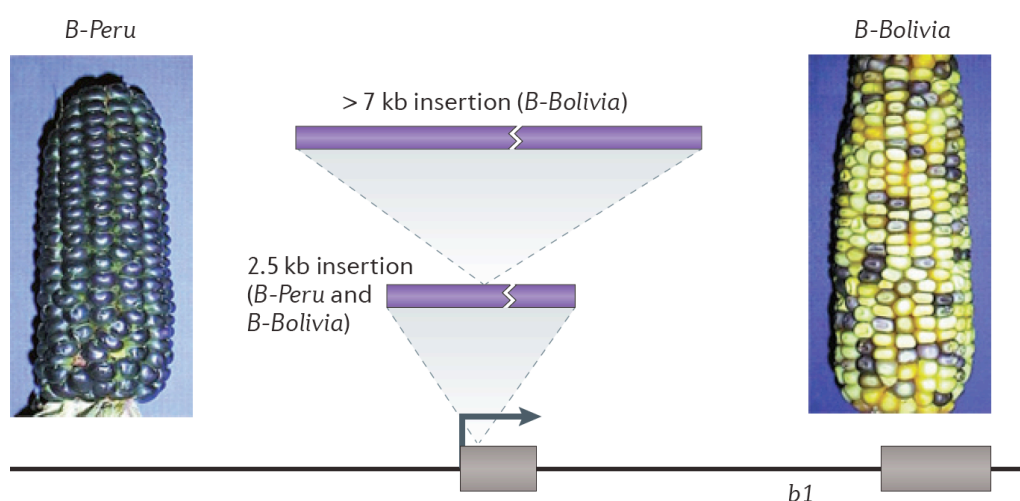


Figure 5. Subsequent Retrotransposons insertions caused gene disruption and phenotypic effects in maize *B-Bolivia* allele (Lisch et al., Nat Rev Genet 2012; reproduced with permission).

Together, those evidences suggest that one reference genome alone is not sufficient to fully represent the genome structure of such a variable genome as the maize one. This led to the proposal of a pan-genome model for maize (Morgante et al., 2007). The pan-genome is a concept that has been first proposed by Tettelin et al., in bacteria (Tettelin et al., 2005) and is composed by a core genome containing sequences that are present in all strains and a dispensable genome composed of partially shared and strain-specific DNA sequence elements.

The pan-genome is probably a concept applicable to several (if not all) plant species. For this reason, SV population studies - coupled with improvements in methods for SV detection - could help to characterize the dispensable genome component, and to reach important achievements in plant genome biology (Marroni et al., 2014).

A recent effort to develop a method in order to build the maize pan-genome, is represented by Pan-genome Atlas (PanA) pipeline (Lu et al., 2015) which exploits both genotyping by sequencing (GBS) and whole genome shotgun (WGS) to produce tag anchors for the construction of the pan-genome: such anchors could be useful to direct and evaluate de novo assemblies of each variety (fig. 6b and c) that will be aligned all together (fig. 6d) and will contribute in the end to the construction of the pan-genome (fig. 6e).

More recently, a maize pan-transcriptome of 368 maize diverse inbred lines has been proposed, and it has confirmed that reference genome is able to capture only half of the maize pan-genes (Jin et al., 2016).

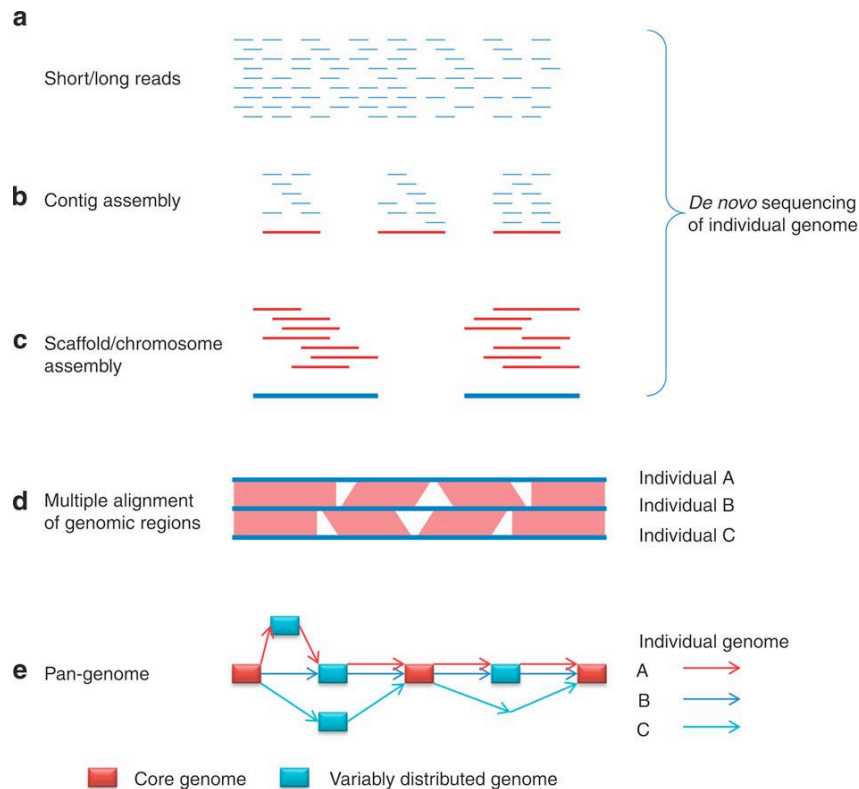


Figure 6. Pan-genome construction through de novo assembly (Lu et al., Nat Comm 2015; reproduced with permission).

The present thesis is part of the NOVABREED project, a European Research Council (ERC) funded project, which aims to characterize the dispensable portion of plants genomes, in order to understand how it contributes to the creation of new genetic variation.

This can be investigated by genome-wide sequencing of different inbred lines of *Vitis vinifera* and *Zea mays*.

An extensive study of composition, origin, function of plants dispensable genome can lead to understanding the genetic diversity source.

Moreover, the project aims to extend such findings to other plant genomes to understand their level of SVs and the mechanisms below them.

1.3 Methods for the detection of structural variation analysis

Although historically an exclusive task performed through CGH arrays, SV analysis is today performed with NGS sequencing (Alkan et al., 2011).

SV originated by TE movements can be detected - comparing sample reads to the reference assembly - as insertions or deletions in the sample relative to the reference.

Several algorithms have been developed in order to detect SVs from NGS sequencing data (Medvedev et al., 2009; Baker et al., 2012).

Such algorithms rely on different approaches, which identify specific SV signatures and patterns:

- Paired-end mapping (PEM) methods search for reads mapping to the reference with an insert size inconsistent with the expected insert size, and/or with anomalous orientation. Deletions are characterized by an insert size longer than the expected (Figure 8); insertions are characterized by pairs in which one read maps close to the insertion breakpoint and the other one (the read generated by the inserted sequence) maps in another genomic location or does not map. The PEM approach is one of the most efficient, in particular for the detection of deletions (Chen et al., 2009), and is often combined with split-read methods (see below) in order to refine the breakpoint coordinates detection (Rausch T et al., 2012; Sindi S et al., 2012). Several approaches have been developed for the detection of insertions with PEM method, (Keane et al., 2012; Hénaff et al., 2015), and all rely on a TE database for the identification of inserted sequences. This prevents the discovery of inserted sequences without homology with known TEs.
- Depth-of-Coverage (DOC) methods are based on the variation of read coverage in a region in comparison to an expected coverage, defined as the coverage in the same region of another genomic dataset used as

control. This method is able to detect deletions, duplications and Copy Number Variants (Abyzov A et al., 2011) over large regions, but the ability of precisely determining the exact borders of the event is limited.

- Split-read (SR) mapping method is able to precisely detect SV breakpoints. The split-read approach requires the use of aligners - such as BWA-MEM (Li H, 2013) and Bowtie (Langmead B et al., 2009) - with split-read mapping capabilities. When a read cannot be mapped entirely, aligners map fractions of the reads separately as a chimeric alignment. The point in which the read is split could be a signature for single-base resolution SV breakpoint detection (Wang et al., 2011; Hart S et al., 2013). In addition, when dealing with insertions supplementary alignments of the split portion of the reads allow locating the origin of the inserted sequence. Extracting information from split read mapping in highly repetitive genomes is a challenging task, where the probability of uniquely mapping a split read is relatively low. However, when this happens, the split read mapping approach offers the unique advantage of enabling the detection of insertions of sequences fragments not annotated as transposable elements (TEs).

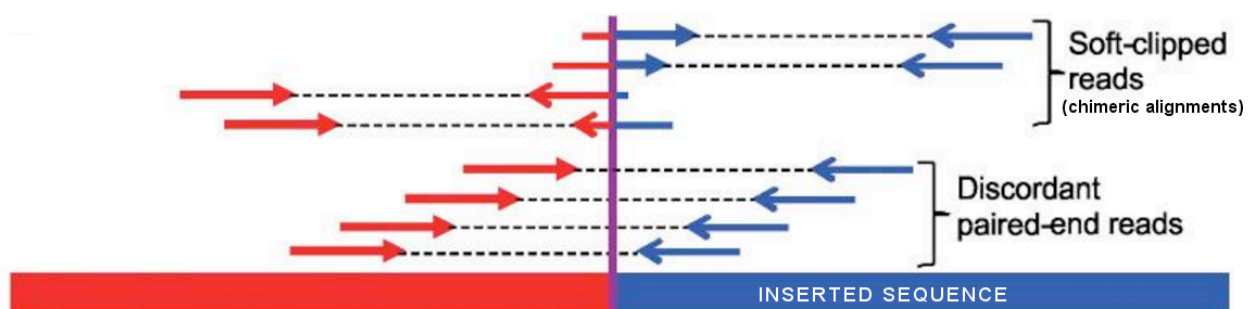


Figure 7. Difference of mapping for discordant paired-end and split reads (soft-clipped reads) on a SV breakpoint (Wang et al., Nat Methods 2011; reproduced with permission).

- De novo assembly methods are potentially able to detect all forms of SV, as it generates contig sequences of the sample that can be then compared to the reference genome. This is in theory a powerful method (Chaisson M et al., 2015), but it is prone to assembly errors in presence of repeated sequences. Both short (i.e. Illumina, www.illumina.com) and long reads (i.e. PacBio, <http://www.pacb.com/>) could be used: the first provide a higher coverage at relatively low costs, while the second allow to resolve highly repeated region, but at comparatively higher costs.

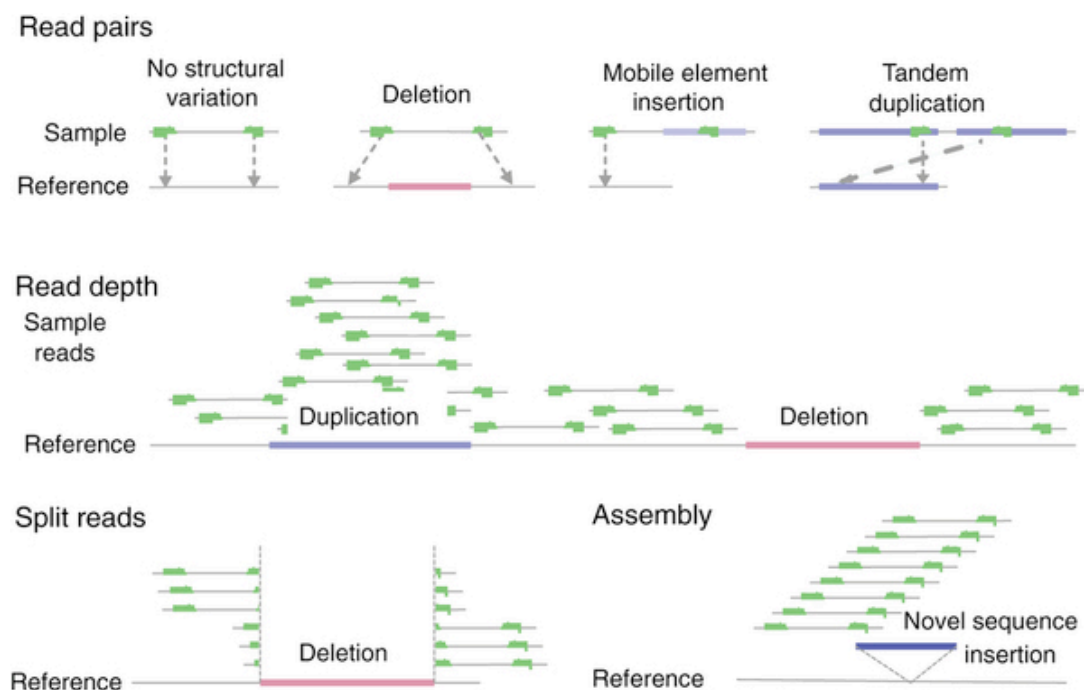


Figure 8. Available methods to find structural variation (Baker, Nat Methods 2012; reproduced with permission).

PEM and SR mapping approaches can be used in combination to increase the ability of SV detection. Delly (Rausch et al., 2012) is one approach that offers an integration between PEM and SR methods to detect more precise deletion breakpoints. Moreover, it is possible to use such methods for the detection of insertions (Hart et al., 2012).

2 OBJECTIVES

The main objective of the present work was the characterization of structural variants (SVs) composing the dispensable fraction of *Zea mays* pan-genome, based on the observation that one reference genome alone is not sufficient to fully represent a single species.

The work has been mainly performed using available tools for the detection of SVs. However, as an additional aim of the work, we developed Walle, a novel algorithm that can detect inserted sequences exploiting split read mapping signatures and without relying on a database of transposable elements. Walle thus fills an existing gap in software tools available for the detection of SVs (and precisely insertions).

We used next-generation sequencing (NGS) data to perform analysis of SVs in genomes of 7 different maize inbred lines, included the B73 reference.

A catalogue of maize structural variants was obtained by using several bioinformatics tools including the aforementioned one.

Since the dispensable portion of the maize pan-genome is mainly composed by transposable elements, we studied their composition and structure in the detected SVs.

Transposable elements may form nested structures so the analysis of them led us to a better understanding of the mechanisms of formation of new structural variations.

The present work resulted in the development of a novel algorithm for the detection of SV and provided insights in the composition of maize pan-genome.

3 MATERIALS AND METHODS

3.1 Library preparation and sequencing

Seven of the eight founder inbred lines multiple-parent advanced-generation inter-cross (MAGIC) maize (MM) population were included in the study: B73 (the reference genome), A632, H99, HP301, F7, W153R, and Mo17.

MM population consists in more than 1,000 maize recombinant inbred lines, 529 of which have been genetically characterized, providing a platform for the study of sequence variants and how they can affect the phenotype (Dell'Acqua M et al., 2015).

Below is an extract of details of the MAGIC maize founder lines we used in our analyses.

Line	Developer	Breeding group	Pedigree
A632	Minnesota Agric Exp Stn	<i>SS</i>	(Mt42 x B14)B14 ³
B73	Iowa Agric & Home Econ Exp Stn	<i>SS</i>	Iowa Stiff Stalk Synthetic C5
F7	INRA-Peronne	<i>Mixed</i>	French population 'Lacaune'
H99	Indiana Agric Exp Stn	<i>NSS</i>	Illinois Synthetic 60C
HP301	Indiana Agric Exp Stn	<i>Popcorn</i>	Supergold
Mo17	Mo Agric Exp Stn	<i>NSS</i>	C187-2 x C103
W153R	Wisconsin Agric Exp Stn	<i>NSS</i>	(Ia153 X W8) Ia153

All the 6 non-reference lines had been previously sequenced in the framework of the project Novabreed.

DNA paired-end libraries were generated from genomic DNA using Illumina Truseq, Illumina TrueSeq PCR-Free, Illumina Nextera (Illumina Inc., San Diego, CA, USA), and Nugen Ovation V2 (NuGEN Technologies Inc., San Carlos, CA, USA) protocols, with a mean insert size of ~400 bp. Libraries were sequenced with

different read lengths on Illumina HiSeq 2500 (2x100 bp and 2x250 bp) and MiSeq (2x300) sequencers, in order to obtain both overlapping and non-overlapping fragment libraries.

Raw data from HiSeq 2500 was processed with the CASAVA 1.8.2 version of the Illumina pipeline while raw data from MiSeq was processed by the instrument integrated MiSeq Reporter Software (MSR) 2.4 version.

Paired end reads from the genomic sequence of Mo17 and F7 were also acquired from the Sequence Read Archive (SRA): SRR764595 (experiment SRX245309), SRR447949 (experiment SRX131286), SRR449556, SRR449557, SRR449558 (experiment SRX132074), SRR1575517 (experiment SRX701271) and SRR1575513 (experiment SRX701267).

Raw sequences were quality trimmed and contaminant filtered using *erne-filter* from *erne tools* version 1.4 (Del Fabbro et al., 2013) and adapters were removed with *cutadapt* version 1.5dev (Martin, 2011).

The resulting trimmed reads underwent Quality Control evaluation using *FastQC* version 11 (Andrews, 2010).

3.2 Alignment and SNP calling

The RefGenV3 sequence of B73 reference genome (Schnable et al., 2009) was obtained from ensemble genomes (Kersey et al., 2012).

Short reads were then mapped against the reference genome sequence using the software package BWA-MEM (Li H, 2013) version 0.7.10 with the -M setting to flag chimeric alignments as secondary.

The Sequence Alignment/Map (SAM) file was sorted and transformed to Binary Alignment/Map (BAM) file with SAMtools version 0.1.19 (Li H et al., 2009).

PCR duplicates were removed with *MarkDuplicates* command of the 1.124 version of the Picard suite (<http://broadinstitute.github.io/picard>) with REMOVE_DUPLICATES=True option, and uniquely aligned reads were selected with samtools quality filter (-q 10).

CollectInsertSizeMetrics tool of Picard suite version 1.88 was used to compute insert size statistics and nt_compute_profile tool of *erne* was used to compute mean coverage.

SNPs were called with Genome Analysis Toolkit (GATK) version 3.2.2 (McKenna A et al., 2010) using paired end read alignments generated by BWA- MEM and filtered with *CleanSam* and *FixMateInformation* of the Picard suite. GATK *RealignerTargetCreator* and *IndelRealigner* routines were used to define intervals in proximity of indels, and to perform local realignment over such intervals. SNPs were called with heterozygosity parameter 0.01 with the *UnifiedGenotyper* routine on each variety. SNPs were retained for further analysis if they were considered of high quality (Phred-scaled quality score > 50) and if they were identified in regions with coverage value not far from the modal value (between 0.5 and 2.5 times the modal coverage value).

3.3 Identification of deletions

For the identification of deletions, two freely available tools were used: DELLY (Rausch et al., 2012) and GASV (Sindi S et al., 2009).

DELLY (version 0.7.2) was used with the default parameters, with alignment files as input. The deletions obtained were filtered for median mapping quality (at least 20), paired-end support (at least 5) and size range (from 1 to 50 Kb).

GASV (version 2.0) was used with the default parameters, preprocessing the alignment files with BAMToGASV software (version 2.0.1).

The obtained deletions were filtered for paired-end support (at least 5) and size range (from 1 to 50 Kb), and discarded when GASV could not explain the data with a single SV (when *Localization* field of the GASV output file value was -1).

For both approaches, when overlapping deletions were found, the one supported by the higher number of PE was retained. Results obtained in each line by the two algorithms were merged in a single dataset. Two deletions were merged in one when both the extremities were closer than 250 bp. Deletions with a number of non-N bases below 1000 were discarded.

Deletions with both extremities closer than 500bp were merged across different lines. This analysis pipeline was developed in our institute and proved to have good performance on simulated and real *Vitis vinifera* data (Gabriele Magris, PhD thesis).

An internally developed Python script (Davide Scaglione, unpublished results) was used to determine the genotype of identified deletions in each sample.

The software analyses how reads from the alignment file map on regions of 500 bp flanking the left and right SV coordinates. Reads were categorized as reads supporting the deletion (or positive reads, with insert size greater than expected, mapping over the detected deletion borders) and reads not supporting the event (or negative reads, mapping inside the deletion with an insert size compatible with the library insert size distribution, supporting the reference genotype).

The genotype was assigned based on the ratio of positive reads to the total number of reads (the sum of positive and negative reads): a ratio below 0.25 was considered homozygous for the reference; a ratio between 0.25 and 0.75 was considered heterozygous; a ratio above 0.75 was considered homozygous for the alternative allele. A semi-supervised script was used to refine overlapping entries: since the genotyping step was run in the whole set of deletions found across varieties, it could happen that deletions not originally identified called by

Delly nor GASV in a specific variety were instead genotyped as heterozygous or homozygous for the alternative allele (ratio at least 0.25).

When such deletions overlapped with deletions identified by Delly or GASV and correctly genotyped, the latter were chosen.

3.4 Transposable elements annotation

A pipeline previously developed by our research group (Gabriele Magris, Andrea Zuccolo and Michele Vidotto, unpublished results) was run in order to annotate detected deletions.

The 1526 sequences of the maize transposable element (TE) database from Wessler et al., version 15-35 (12th February 2015) together with 2000 sequences obtained from RepBase (Jurka J et al., 2005) were used for the annotation.

Annotation was performed in 5 incremental steps; in each step, only items not identified in the previous steps were analyzed:

- Tandem Repeats Finder (TRF) (Benson G, 1999) was run with the option pattern site < 170 bp (the period size of repeats to consider) in order to mask tandem repeats regions and the resulting masked regions for more than 80% of the sequences were excluded from further analysis.
- RepeatMasker (Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open-3.0*) was used to compare both edges (400 bp) of each deletion with edges (100 bp) of each TE in the database, in order to associate a TE superfamily to related deletions. At this step, a deletion was associated to a TE superfamily if both its edges were associated to the same superfamily.
- RepeatMasker was used again in order to identify LINE or solo-LTR elements, when an 80% match with a database element was found.
- For the discovery of unannotated LTR elements LTR_Finder (Xu Z et al., 2007) was used and a deletion was associated to an LTR element if both

its edges (400 bp) were identified as putative LTR regions.

- REPET Teannot package (Flutre T et al., 2011) was run to classify not annotated deletions so far, in comparison to previously annotated elements: this was done by aligning all the deletions to the maize TE database with different algorithms (Blaster, RepeatMasker and CENSOR); all high-score pairings (HSP) found are filtered and concatenated by MATCHER; distant fragments were connected with the "long join" procedure, in order to obtain final annotations.

3.5 Dating of Long Terminal Repeat insertion events

Sequence divergence between complete paired Long Terminal Repeats (LTRs) can be used to date LTR retrotransposon insertions (SanMiguel P et al., 1998). A pipeline for dating of LTRs previously developed by our group was used (Gabriele Magris, Andrea Zuccolo and Michele Vidotto, unpublished results), and is described below. Precise coordinates of 5' and 3' LTR were obtained using LTR_FINDER on the reference genome with the following parameters: D=50000 (Max distance between LTRs), d=100 (Min distance between LTRs), L=6000 (Max LTR Length), l=50 (Min LTR Length), p=15 (length of exact match pairs), E (LTR must have edge signal), C (automask highly repeated regions), s (predict PBS by using a tRNA database), a (use ps_scan to predict protein domain), F=11110000000 (results must have both 5' and 3' LTR with TG and CA signatures).

LTRs were separately identified genome-wide from retrotransposons in regions not involved in SV in our maize varieties, and within deletions ranges (extended by 400bp at both sides to adjust for possible imprecision in deletion breakpoint estimation), considering only LTRs found at less than 500bp from the breakpoint coordinates.

Couples of LTR sequences were pairwise aligned with the EMBOSS Stretcher aligner (Rice et al., 2000) in order to compute the evolutionary distance (K) between each couple, using EMBOSS distmat tool with the Kimura's Two-Parameter method (Kimura, 1980).

The time of insertion was calculated for each retrotransposon via the equation

$$T=K/2*k$$

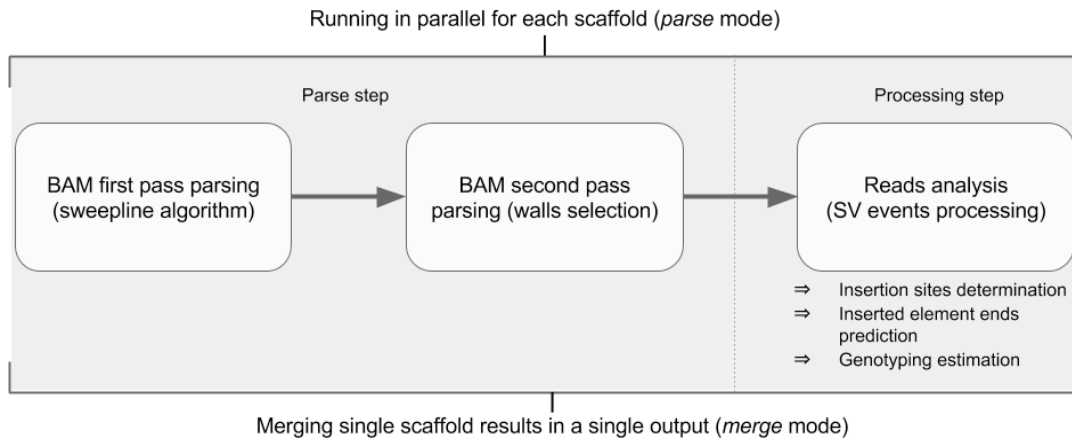
where K is the evolutionary distance and k is the substitution rate of 1.3E-08 mutations per site per year (Ma et al., 2004), 2-fold higher than the synonymous substitution rate previously observed for the adh1 and adh2 loci of grasses (Gaut et al., 1996).

3.6 Walle algorithm

Walle is a Python software for the detection of structural variants (SVs), developed in the present project.

The main focus of Walle is the detection of insertions using split read alignments and using stacks of split reads (informally called “walls”, hence the name) for the precise detection of insertion breakpoints. Compared to existing software for the detection of insertions, Walle does not rely on any transposable elements database. The only input required is a BWA MEM alignment .bam file and its .bai index file.

The algorithm (*parse* mode) proceeds in two main steps, the parsing step and the processing step. Both steps can be run in parallel on each scaffold separately to reduce computation time when multi-core systems are available. A *merge* mode is implemented to put together each scaffold result in a single output.



During the parsing step, the alignment file is read through a sweep line algorithm implementation (Shamos et al., 1976), looking for stacks of split reads, or “walls”: those stacks are the results of chimeric alignments around a SV breakpoint, described in the SAM format specifications as “an alignment of a read that cannot be represented as a linear alignment”.

In a chimeric alignment, a read is split in two or more parts, each mapping to different positions of the reference genome, since the aligner splits the sequence read in subsequences (soft-clipping) with the same query template name (QNAME) and possibly marks one as *representative* (usually the longer subsequence) and others as *supplementary* alignments (Fig. 9).

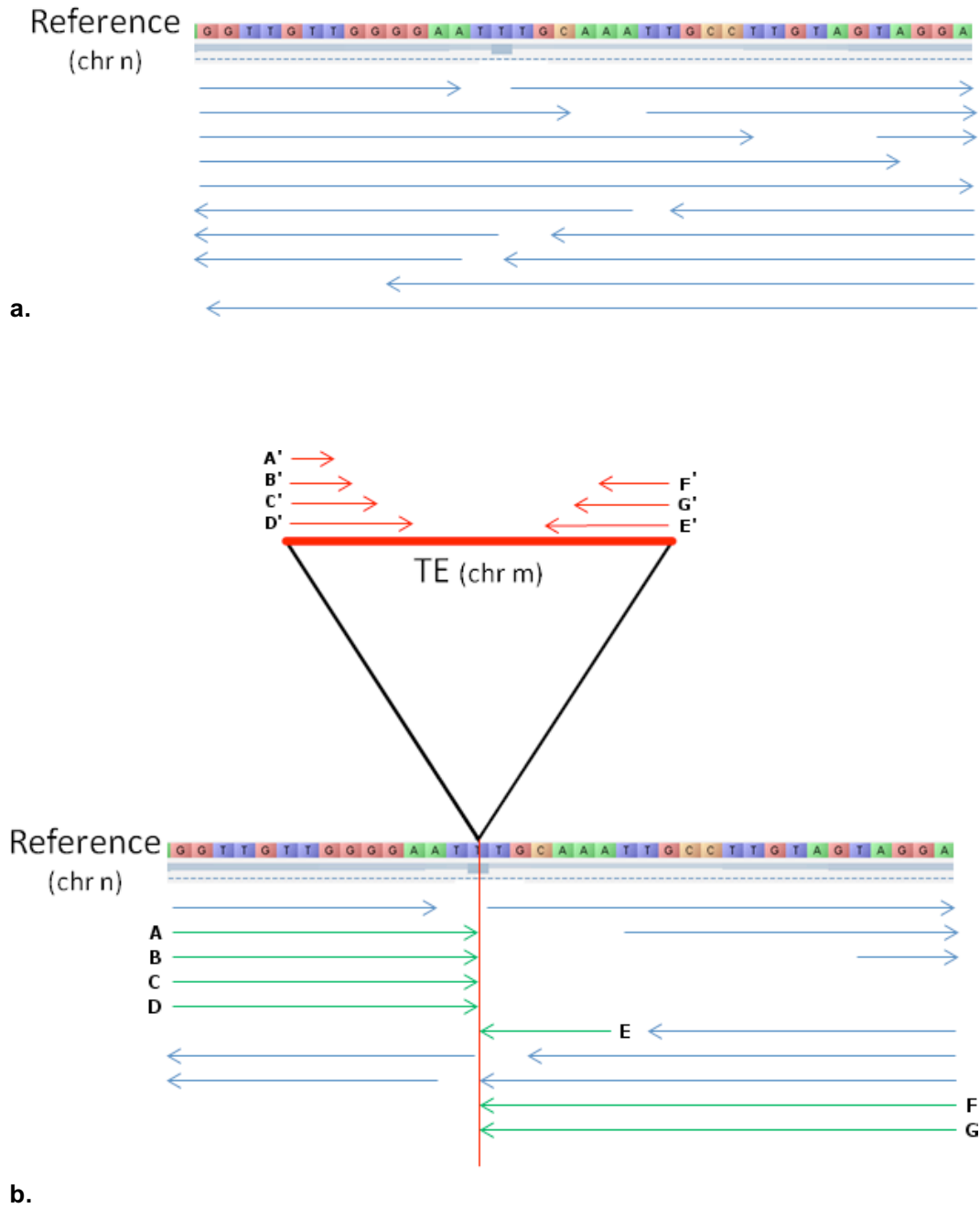


Figure 9. An example of linear (a) and chimeric (b) alignments.

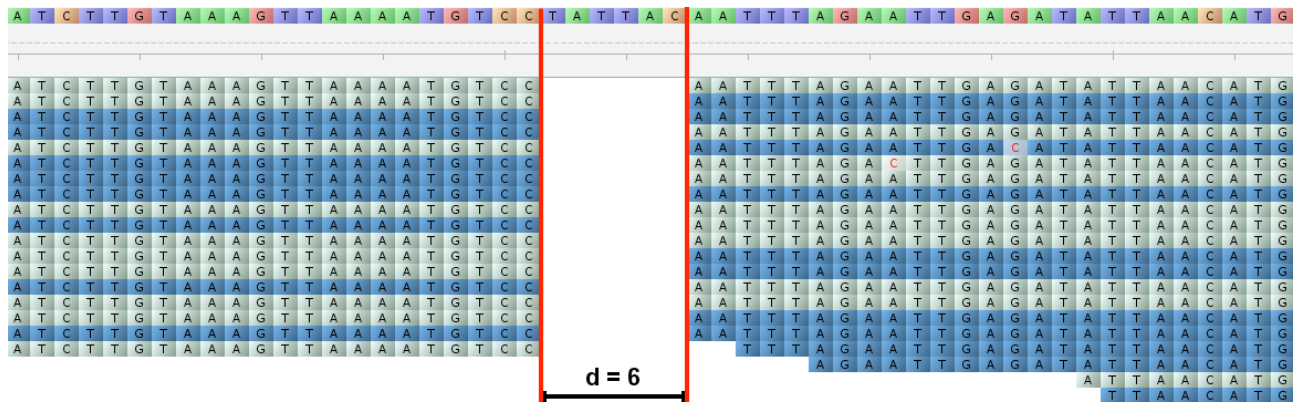
In regions with linear alignments (a), each read maps as a unique sequence, without any splitting. In regions with chimeric alignments (b), reads split in 2 (or more) sequences which can map in different loci of the reference genome; this usually happens across breakpoints due to structural variations, such as an insertion of a transposable element (TE). Exploiting information of split read mapping, it is possible to detect insertion sites (red vertical line), farther to infer inserted element reference coordinates. In the example (b), reads colored in green are split and oriented towards the breakpoint end (A,B,C,D) and start (E,F,G) around the same reference coordinate, which is the putative breakpoint coordinate itself (red vertical line); each split read could have a supplementary alignment and, in the case of a TE insertion, supplementary reads (colored in red) map at the beginning (A', B', C', D') or at the end (E', F', G') of the inserted element, depending on where their respective primary read maps.

A second pass of parsing is done around the detected breakpoints in order to collect reads details and to infer the type of event generating the chimeric alignments later in the processing step. Regions with high local coverage (default is *mean coverage* * 3) can be skipped to increase software performance.

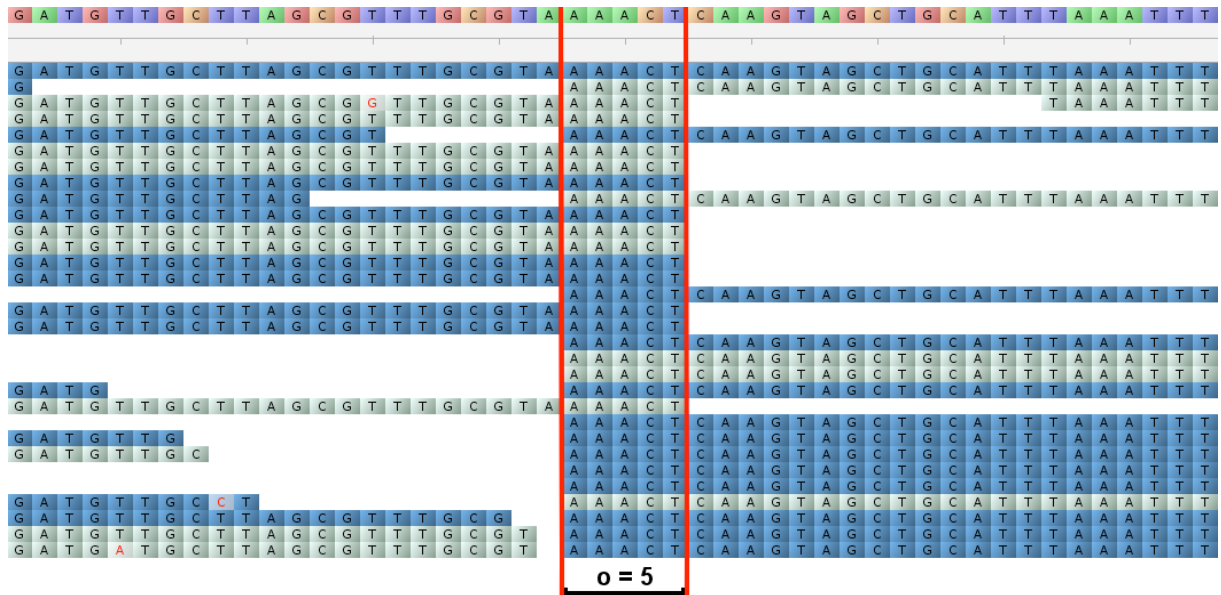
In the presence of an event, such as an insertion, split reads are distributed in two stacks on both sides of the insertion breakpoint in the reference genome (Fig. 9b), and Walle discriminates left from right “walls” during the parsing step.

Left and right “walls” are then coupled as putative SV breakpoint events by proximity and abundance, depending on user-defined parameters (Fig. 10): *distance* (-d, the maximum distance between 2 stacks of reads to call an event), *overlap* (-o, the maximum overlap permitted between 2 stacks to call an event) and *cutoff* (-c, the minimum dimension of a split-reads stack, or number of reads composing a “wall”).

A reparsing mode is implemented to perform the second pass only, giving a list of candidate breakpoints as additional input (*reparse* mode).



a.



b.

Figure 10. Walle proximity parameters examples, from two different breakpoints visualized on Tablet (Milne et al., 2012).

Option $-d$ described in (a) is the maximum allowed distance at which they can be two stacks of split reads, ($d = 6$ in the example).

Option $-o$ described in (b) is the maximum allowed overlap at which they can be two stacks of split reads ($o = 5$ in the example).

Reads are colored by orientation: forward reads in pale green, reverse reads in blue. Abundance is evaluated by the number of split and corrected oriented reads in a single stack: for example, in (a), 10 reads contribute to the left stack, while 8 reads contribute to the right stack; the breakpoint will be predicted only for $c < 8$.

In the processing step each event is then analyzed to infer SV type and zygosity. An insertion breakpoint is defined when most of the reads have supplementary alignments or mates mapping on another chromosome or on the same chromosome at a distance greater than a given value (200 Kb is the default) on both sides of the breakpoints.

This detection rule is based on the evidence that reads spanning an insertion breakpoint can be mapped as split reads with a primary alignment (Figure 11, L1 and R1) flanking the breakpoint and their respective supplementary alignment (L1' and R1') mapping on the edge of the inserted element (i.e. a transposable element). While primary alignment reads have an orientation toward the breakpoint, their mate reads (L2 and R2) will also map on the inserted element, otherwise they will map externally to the insertion. In the first case, they will map on an inner position of the inserted element, with respect to the secondary alignments.

The inserted sequence position in the reference genome is then estimated by looking where such supplementary alignments map, when present, or where mate reads map.

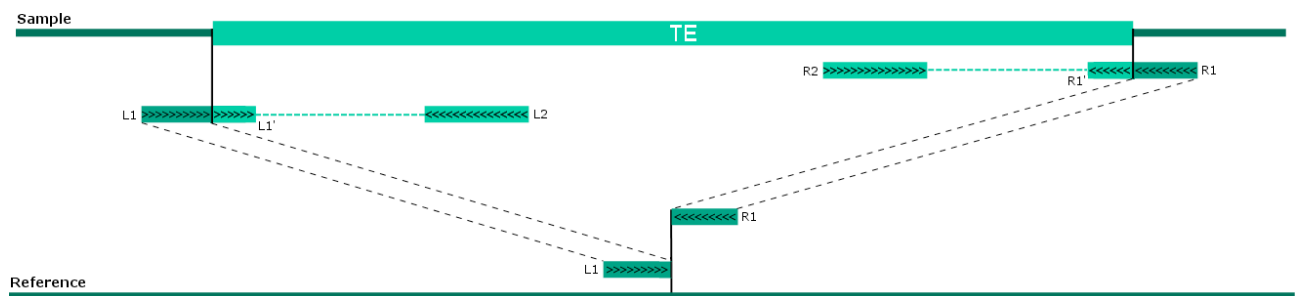


Figure 11. Chimeric alignment and pair mapping in the presence of a TE insertion.

Reads flanking an insertion breakpoint are split (L1, R1), possibly with a supplementary alignment on each TE end (L1', L2'). While correctly oriented toward the breakpoint (inner black arrows), they also have mates mapping inside the TE element (L2, R2) in another chromosome, or with a much greater insert size than the mean of the library.

A measure of uncertainty of the estimated coordinates has been developed as follows. First, a score is attributed based on the number of supplementary alignments supporting the position. With a score of 2, both the inserted element coordinates are estimated by supplementary alignments, while a lower score, indicates that one (score = 1) or both (score = 0) the coordinates are defined by mate reads mapping. It is important to note that in those cases the dimension of the inserted element itself is probably underestimated, as mate reads should map more internally on the element, with an error that can be conservatively estimated as:

$$(\text{mean insert size} - \text{read length}) * (2 - \text{score})$$

hence proportional to the number of split-reads supporting the definition of edges of the inserted region (Fig. 12).

The estimation of inserted element coordinates can be re-run separately given a list of insertion breakpoints as additional input (*find_insert* mode).

The ratio between split and non-split reads flanking the breakpoint is computed for each call and used for inferring the call quality.

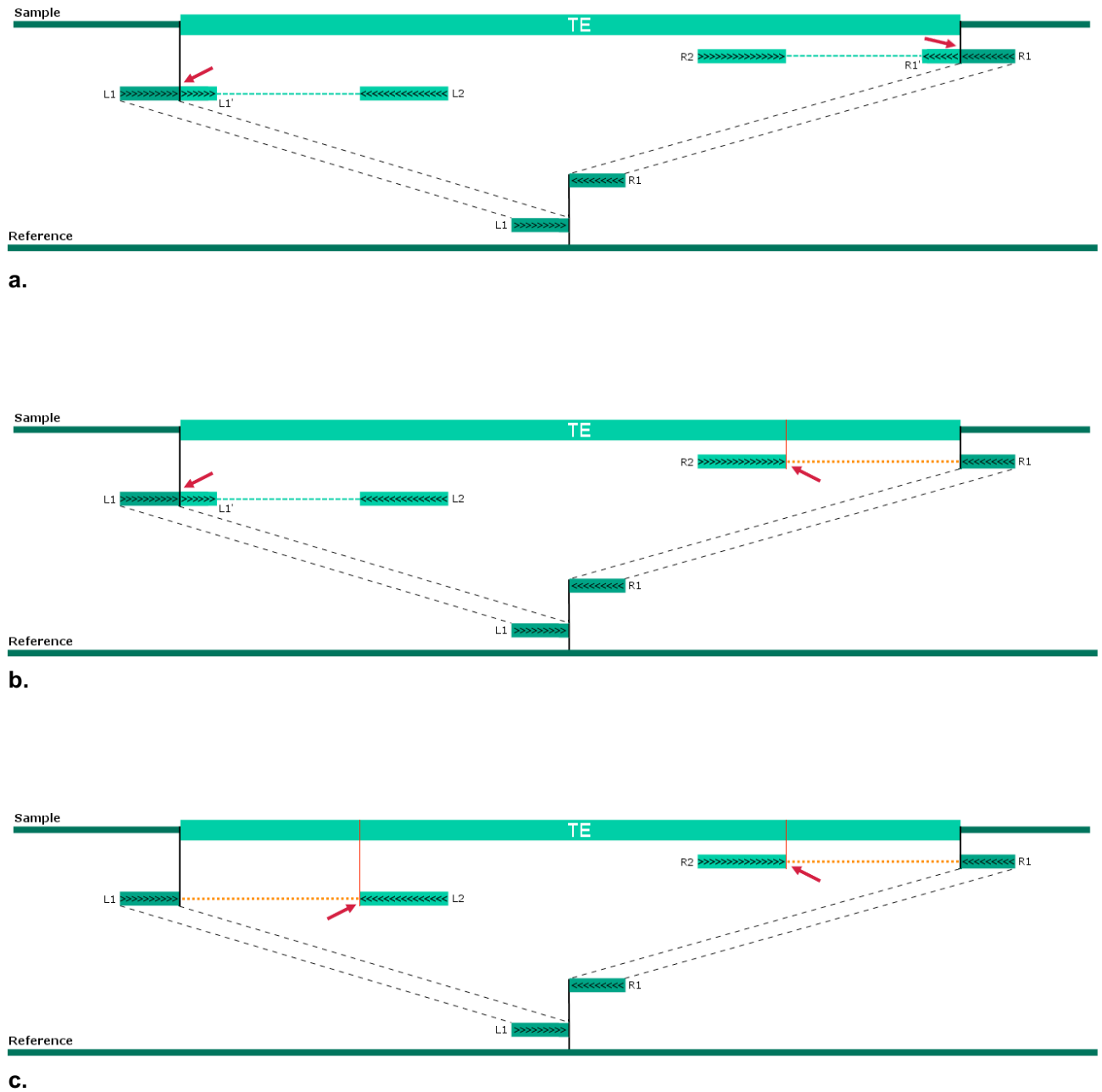


Figure 12. Reference coordinates of the inserted element definition and score explanation.

When supplementary alignments are found at both ends of the inserted elements (a), the predicted coordinates report a score of 2 (number of ends supported by supplementary alignments), and the uncertainty in their definition is null. When only one end of the inserted elements is supported (b), the reported score is 1 and the other end is defined by mate information, so the error in the definition of that end is proportional to the distance of the mate read from the end itself, which is similar to the mean insert size of the library. When no ends of the inserted element are supported by supplementary alignment (c), the reported score is 0 and both ends are defined by mate information, so their error is similar to twice the mean insert size of the library. Red arrows indicate the predicted ends; black vertical lines indicate the real ends; red lines indicate imprecise predicted ends; orange dashed lines indicate errors between predicted and real ends.

Walle has been developed in Python, can be freely downloaded from <https://bitbucket.org/ezapparoli/walle/> and is fully documented. We report below an extract of the *help* page (*-h* option). Required arguments are in bold:

```

-f, --file           the input BWA-MEM bam alignment file (required)
-m, --mode          mode of running, could be: parse, merge, reparse, find_insert (required)
-s, --seq           the contig to process (required)
-c, --cutoff        the minimum dimension of a split reads stack (required)
-d, --distance      the maximum distance between 2 stacks to call an event (required)
-o, --overlap       the maximum overlap permitted between 2 stacks to call an event (default is 5)
-i, --isize         the mean insert size of the library (estimated if not given)
-e, --expected_coverage expected mean coverage of the library (estimated if not given)
-b, --bam           write bam file of split reads found
-C, --coords_file   coords file input for the reparse and find_insert step
-v, --verbose       verbose actions

```

3.7 Detection of insertions on simulated data

For the detection of insertions, we tested the performance of an internal pipeline developed in our group (Sara Pinosio, PhD thesis) and already used for *Populus trichocarpa* (Pinosio et al., 2016) and *Vitis vinifera* TE insertions detection (Gabriele Magris, unpublished results), of the newly developed Walle and of two freely available tools: Jitterbug (Hénaff et al., 2015) and Retroseq (Keane et al., 2013).

The pipeline developed by Sara Pinosio (Pinosio et al., 2016) attempts to identify insertion breakpoints by looking at groups of reads mapping as singletons; reads flanking the insertion site are de novo assembled using CAP3 (Huang X. et al., 1999) for each side and each resulting sequence is mapped on the reference genome: a putative insertion breakpoint is confirmed if the reconstructed

sequence maps with opposite orientation and at distance lower than the mean sequenced library insert size. The unmapped mates of reads mapping as singletons are de novo assembled using CAP3 in order to reconstruct two consensus sequences of each inserted element edge region. Each couple of consensus sequences is aligned to a TE database using blastn and if a couple aligns to the same TE extremities, the insertion is called.

The tools were tested on alignments of real reads on a simulated reference of *Vitis vinifera*, previously developed by our group (Sara Pinosio and Gabriele Magris, unpublished results). A similar benchmark in a maize simulated genome was not possible, as public paired-end libraries with adequate insert size distribution were not available for B73.

In the simulated reference, 1000 repeated sequences, ranging in size between 1 and 25 Kb, were randomly moved from their original positions in the genome and inserted to randomly chosen new positions, simulating 1000 TE movements. The alignment was produced with BWA MEM, with default parameters.

The internal pipeline, Jitterbug (version 1.0, personal communication with Elizabeth Hénaff) and RetroSeq (version 1.5) were used with default parameters. For the internal pipeline, we considered calls with more than 5 PE supporting the event. For Jitterbug, both the default filter as described in Hénaff et al., and a more relaxed filter were used; to summarize both the filters settings we apply the following cutoffs:

Jitterbug filters cutoffs	<i>Default filter</i>	<i>Relaxed filter</i>
<i>Max cluster size</i>	5 * coverage	15 * coverage
<i>Max span</i>	mean_fragment_length	3 * mean_fragment_length
<i>Max clipped support</i>	5 * coverage	15 * coverage
<i>Min interval length</i>	mean_fragment_length	0.5 * mean_fragment_length

According to Jitterbug paper: max cluster size is the maximum number of reads for a cluster to be considered; max span is the maximum distance allowed between two reads start positions in a cluster (similar to --distance concept of Walle algorithm); max clipped support is the maximum number of clipped reads supporting the insertion position; min interval length is the minimum length of the predicted insertion interval.

For Retroseq, only calls passing all internal filters (FL=8) and with a genotype quality (GT) greater or equal than the first quartile of the GT distribution across all the calls with FL=8 were considered.

Walle was used with cutoff of at least 3 split reads to call a breakpoint, at maximum distance of 20 bp and with a maximum overlap of 10 bp.

All the tools with the exception of Walle rely on a transposable elements database; in this case the database was composed by the 1000 deleted sequences. Results were evaluated in terms of insertions called by each tool compared to the set of simulated insertions, reporting true positive rate and positive predictive value. Non-simulated positives calls were used as an approximation of false positives (FP), since they are the sum of false positives and real insertions in the reads compared to the assembled reference genome. True positive (TP) calls are calls which are simulated insertions, while false negative (FN) calls are simulated insertions which are not called. The sum of TP and FN is always 1000, which is the total of simulated insertions. In other terms, the null hypothesis is when a position of the genome is not a simulated insertion breakpoint and is not called as it; a TP corresponds to rejecting the null hypothesis, that is a call of a simulated insertion breakpoint.

3.8 Detection of insertions in maize

Walle and the pipeline previously developed by our group (Pinosio et al., 2016) were used for the identification of insertions in MAGIC maize population. This combination was chosen because of the good performances and the

complementary nature of the two methodologies: Walle allows precise breakpoint detection and detection of insertions not annotated in the databases, the pipeline is able to fully exploit TE db for the identification of annotated transposable elements.

Bam files obtained aligning paired reads of the 6 maize lines to the reference using BWA-MEM were used as input for both detection tools.

As TE database, we used the Maize transposable element (TE) database from Wessler et al., version 15-35 (12th February 2015), including deletion sequences identified by our analysis with DELLY and GASV, in order to maximize the discovery of insertions to annotated and not annotated transposable elements.

To improve Walle detection, a reparsing on insertion breakpoints identified in the previous step was performed using longer reads obtained from non-overlapping fragment libraries. The greater length of reads increases the probability that split reads are mapped unambiguously to the reference genome, and therefore the chances of correctly identifying the inserted sequence.

Insertions detected by the TE db-dependent algorithm were merged as single events across varieties if breakpoints were less than 250bp apart from each other.

Insertions detected by Walle were merged as single events across varieties when two or more breakpoints were overlapping.

Genotypic status of insertions detected by Walle was determined by a routine of Walle. Since the internal pipeline for the detection of insertions is not able to detect genotyping status, we estimated genotypic status of insertions using a Python script (Davide Scaglione, unpublished results) similar to one used for deletions.

The script searches for positive reads, defined as reads that are located less than 500bp apart from the breakpoint, and have a mate a) aligned to another genomic position and b) mapping on 5' or 3' of the inserted transposable element; while negative reads are reads supporting the reference genotype, with an insert size not different to the expected one (similar to the mean insert size of the library).

Similarly to what has been done for deletions, the genotype was inferred by the ratio of positive reads to the total number of reads (the sum of positive and negative reads): a ratio below 0.25 was considered homozygous for the reference; a ratio between 0.25 and 0.75 was considered heterozygous; a ratio above 0.75 was considered homozygous for the alternate allele.

Results from each of the two tools were integrated in a single dataset by merging insertion events when 1-bp overlap of breakpoint intervals were found and the larger breakpoint was reported.

3.9 Analysis of genes affected by SV

Genes interrupted by an insertion or which undergo a deletion were identified and classified on the basis of the evidence that the variant occurs on an exon, an intron, or both. To functionally characterize genes affected by SV in at least one exon, we performed functional annotation according to gene ontology.

Zea mays RefGen_v3 annotation build (5b+) and optimized ontology was retrieved from Gramene archive (Tello-Ruiz et al., 2016).

Overrepresentation of GO terms in genes affected by SVs were tested by Fisher exact test integrated in topGO (Aibar et al., 2015).

Only GO terms with a Fisher exact test p-value < 0.05, not corrected for multiple testing, were considered.

3.10 Validation of SVs on *de novo* assembly

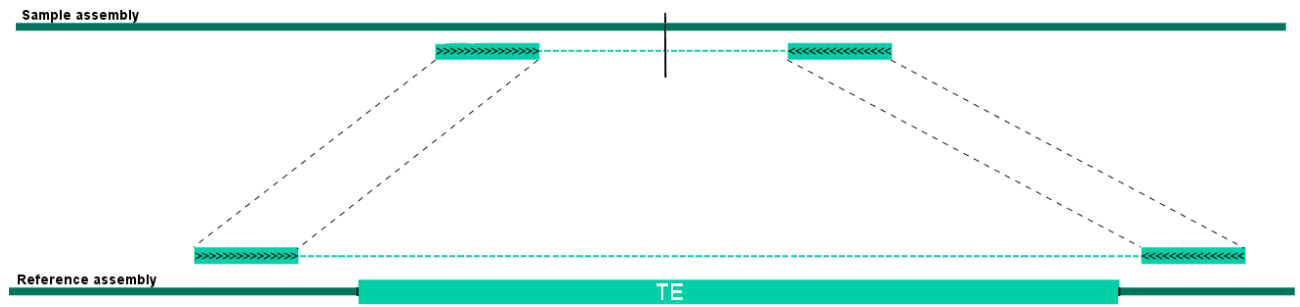
Structural variants were validated using *de novo* assemblies obtained with ALLPATHS-LG (Gnerre et al., 2011) for A632, H99 and HP301 varieties (Michele Vidotto, unpublished data).

For each line, reads supporting each SV event when mapping on the B73 reference were mapped on the *de novo* assembly of the line itself.

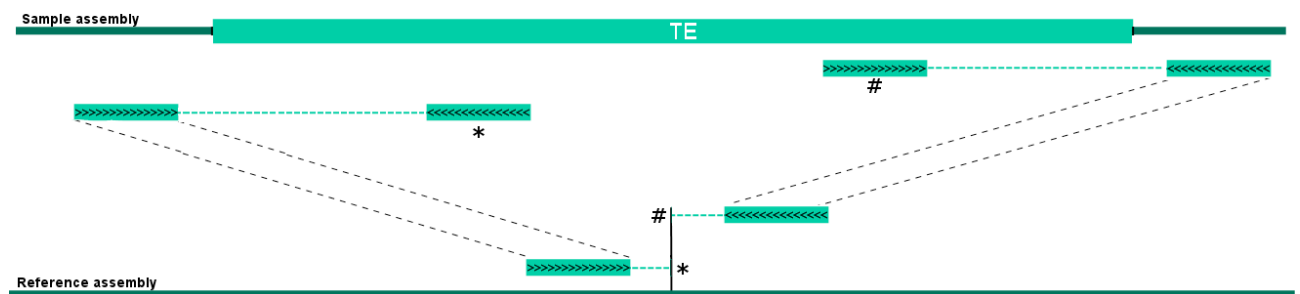
The expectation is that the assembled contigs do not contain the sequence that is present in B73, and that such reads should map as a proper pair on the assembled contigs.

For each SV, all the reads supporting the variant itself on the B73 reference were extracted; such reads will be referred to as “positive” reads. Alignment of positive reads on the assembled contigs was examined; their insert sizes were evaluated, grouped by each SV event. SV events were evaluated when at least 5 positive reads were found to align to the assembly.

A true positive deletion is defined when all positive reads mapping to the same contig map in a proper pair, i.e. they have an insert size similar to the library mean insert size, and are correctly oriented (Fig. 13a). Conversely, a false positive deletion is defined when at least one positive read pair is mapping not as a proper pair, i.e. with an insert size greater than the library mean insert size. Similar criteria have been adopted for insertions. In this case, the expectation is that the assembly contains the inserted sequence (absent in B73) and that reads that were mapped as singletons and/or as split reads on B73, will now map as a proper pair (Fig. 13b). True positive insertions are defined when all positive reads map to the assembly in a proper pair, while a false positive insertion is defined when at least one positive read pair is mapping not as a proper pair. Reads at either side of the breakpoint can be used for validation.



a.



b.

Figure 13. Reads supporting a deletion (a) and an insertion (b) in respect to the reference assembly, maps properly in the sample assembly from which they were sequenced. Thus, a SV is validated if its supporting reads map properly in the sample assembly.

4 RESULTS AND DISCUSSION

4.1 Sequencing

Six of the seven MAGIC maize (MM) parental lines analyzed in this study were sequenced with different technologies to obtain two sets of read lengths, defined as non-overlapping and overlapping reads, respectively. For F7 and Mo17 lines we have also downloaded available short non-overlapping reads from the Sequence Read Archive (SRA). The seventh line was the B73 reference which is publicly available as an assembled genome.

A summary of metrics of the resulting set of reads are reported in tables 1a and 1b for non-overlapping and overlapping reads, respectively. In Table 1a, read lengths of all sequenced libraries are listed, as they were partly downloaded from SRA. In Table 1b, a mean of read lengths of all sequenced libraries is reported, for each line.

Table 1a: Sequencing results for non-overlapping reads

Line ¹	Coverage ²	% Ref covered ³	Physical coverage ⁴	Read length ^{5a}	Mean Insert size ⁶
A632	17.71	63.81	24.97	100	402
H99	13.47	48.51	24.01	100	377.28
HP301	16.59	60.61	25.51	100/125	399
F7	31.99	57.86	45.97	100	404.34
W153R	10.72	59.81	18.64	100/125	412.5
Mo17	23.07	52.33	36.36	75/90/100	250.53

¹ Name of the *Z. mays* line. ² Mean sequence coverage of the uniquely aligned reads. ³ % B73 reference genome covered by unique aligned reads. ⁴ Mean physical coverage of the unique mapped reads. ^{5a} List of read lengths. ⁶ Mean library insert size (bp).

Table 1b: Sequencing results for overlapping reads

Line ¹	Coverage ²	% Ref covered ³	Physical coverage ⁴	Read length ^{5b}	Mean Insert size ⁶
A632	27.16	72.05	40.5	268.75	417.12
H99	29.31	65.35	60.99	284.37	396.3
HP301	28.12	69.05	49.36	283.33	382.73
F7	23.27	67.73	40.37	268.75	415.65
W153R	25.26	69.3	42.28	265	391.25
Mo17	24.24	68.95	39.78	266.66	445.6

¹ Name of the *Z. mays* line. ² Mean sequence coverage of the uniquely aligned reads. ³ % B73 reference genome covered by unique aligned reads. ⁴ Mean physical coverage of the unique mapped reads. ^{5b} Mean read lengths. ⁶ Mean library insert size (bp).

For non-overlapping reads, a resulting mean coverage of approximately 19X was obtained, with a standard deviation with 7.64. Downloaded reads have contributed to raise coverage for F7 (32X) and Mo17 (23X), while W153R remains the lowest covered line (less than 11X).

Overlapping libraries have higher coverage with a mean coverage of 26X and a standard deviation of 2.34.

H99 covered the lowest proportion of B73 reference genome, while A632 covered the highest proportion. This is true for both non-overlapping and overlapping reads datasets, although the latter have higher proportions in general, as expected with higher coverage and for longer reads.

Both read lengths and insert sizes are generally uniform across lines in the two sets of sequencing data, except for Mo17 downloaded short non-overlapping reads, which have shorter reads and smaller insert sizes.

4.2 Software development for the identification of insertions

Walle, an algorithm for the identification of insertion sites using paired-end and split reads alignments has been developed.

The major difference compared to existing insertion discovery software is the ability to predict insertions without the use of a TE database.

In addition, Walle is the first software taking full advantage of split read mapping, while existing software do not make use of split-read mapping or use it only for the refinement of breakpoints. Walle insertion breakpoints discovery relies on split reads, while paired-end information is used in a later step in order to allow the prediction even when the secondary alignment mapping component is low. That may be due to high repetitive sequences of TE edges, i.e. terminal-inverted repeats or LTR sequences.

Our algorithm exploits the evidence that in the case of an insertion in a sample compared to the reference, the alignment of reads of that sample on the reference tend to form a structure similar to a wall which coincides with the insertion breakpoint, as reads are split at the exact coordinate where the breakpoint itself occurs.

To evaluate the effectiveness of Walle, we tested it on simulated data, and compared performance with that of similar tools.

Walle, Jitterbug (Hénaff et al., 2015), Retroseq (Keane et al., 2013), and another pipeline developed by our group (Pinosio et al., 2016) were run separately on alignments of real reads on simulated *V. vinifera* genome, where 1000 sequences were inserted at random positions. We then integrated results of the 2 best performing algorithms, in order to maximize their performances, and used this integration as the best performing approach for the detection of real insertions in *Zea mays*. We will refer to this approach as “Merged”.

As reported in Table 2 and in Figure 14, Walle is the method with the higher balance between power (or sensitivity, True Positives rate) and specificity (True

Negative rate), with a F1 score of 0.91; Retroseq has the highest power (0.93) but it suffers a high number of False Positive calls, which drop the F1 score to 0.85; the Pinosio et al. pipeline is the one with the best performance after Walle, with a F1 score of 0.87, hence it was chosen for integration with Walle results. The merged dataset test results in a F1 score of 0.95, with a sensitivity of 0.97 and 80 False Positives calls.

Table 2: Benchmarking of tools for the detection of insertions

Tool ¹	TP ²	FP ³	FN ⁴	F1 ⁵	Power ⁶
Pinosio et al.	808	50	192	0.87	0.81
Jitterbug	745	52	255	0.83	0.74
Retroseq	929	248	71	0.85	0.93
Walle	857	36	143	0.91	0.86
Merged	972	80	28	0.95	0.97

¹ Name of the tool tested; Merged is the integration of Walle and Pinosio et al. tool.

² True Positives. ³ False Positives. ⁴ True Negatives. ⁵ Power or sensitivity or True Positive rate.

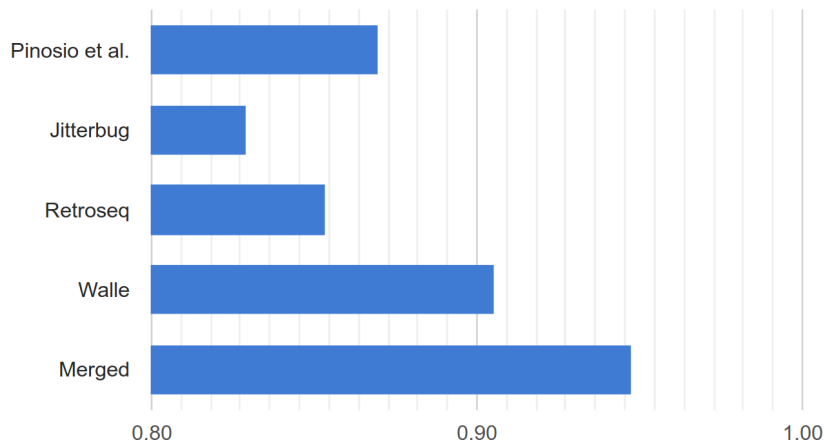


Figure 14. F1 score comparison of different algorithms tested, including the merged dataset. “Pinosio et al.” is the pipeline developed by University of Udine and IGA (Pinosio et al., Mol Biol Evol 2016), Jitterbug and Retroseq are two publicly available software packages for SV detection, Walle is the tool developed in this work and Merged is the dataset resulting from the integration of Walle and Pinosio et al. algorithms. Walle is the single best performing tool, followed by Pinosio et al. Merging of the two further increases F1.

Since Walle breakpoint prediction is based on split reads signatures, which are likely more precise than paired-end mapping signatures, we expect that breakpoint definition performed by Walle is more precise than that performed by other methods. We tested this by comparing the distribution of the distance of the simulated insertion from the breakpoints defined by each tool (Figure 15). As expected, Walle has the best overall results, with a median breakpoint error of 3 bp. Jitterbug outperforms Retroseq with the same median distance of 7 bp, but a tighter distribution, with a standard deviation of respectively 2.1 bp and 17.6 bp. The merged results showed a slightly wider distribution of distance from the simulated insertion compared to the results obtained by Walle alone, because in the merged data, several calls were not supported by split reads (i.e. were private of the pipeline by Pinosio et al.) and had on average higher distance from the simulated insertion breakpoint.

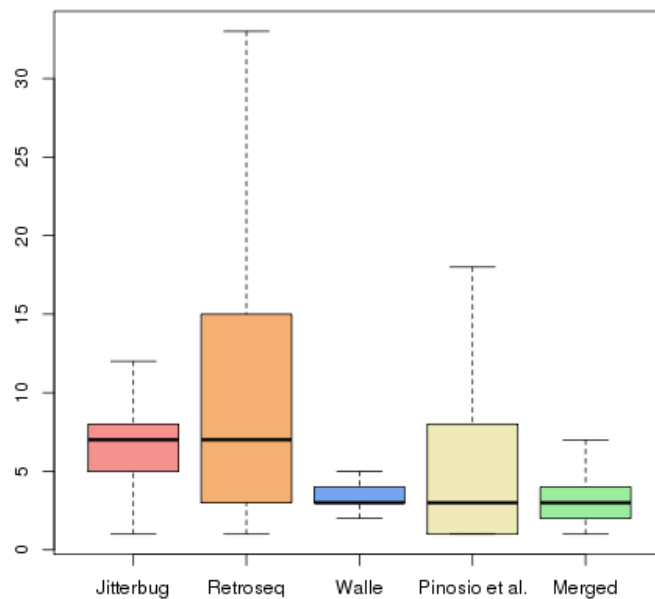


Figure 15. Distributions of distances from true breakpoint of each tool, including the merged dataset.

“Pinosio et al.” is the pipeline developed by University of Udine and IGA (Pinosio et al., Mol Biol Evol 2016), Jitterbug and Retroseq are two publicly available software packages for SV detection, Walle is the tool developed in this work and Merged is the dataset resulting from the integration of Walle and Pinosio et al. algorithms. While Walle (blue) has the narrowest distribution among other algorithms, its median is the same of Pinosio et al. algorithm (yellow). Merged (green) is the dataset resulting from the integration of the two algorithms, which has the same median, too.

In figure 16 we show sensitivity of each tool assuming that a call is retained as true if the distance from the real event is equal or lower to X bp, where X is varying from 1bp to 10 bp.

While Walle reaches a sensitivity greater than 0.75 at 5 bp, other tools don't reach such sensitivity levels within a 10 bp distance from true breakpoints. The only exception is the merged dataset, which reaches a sensitivity of 0.75 at 8 bp.

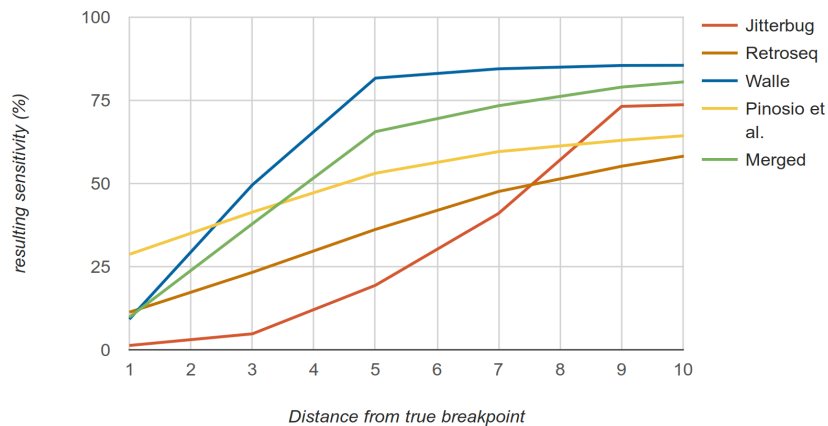


Figure 16. Accuracy of breakpoints predicted by each tool, including the merged dataset.

“Pinosio et al.” is the pipeline developed by University of Udine and IGA (Pinosio et al., *Mol Biol Evol* 2016), Jitterbug and Retroseq are two publicly available software packages for SV detection, Walle is the tool developed in this work and Merged is the dataset resulting from the integration of Walle and Pinosio et al. algorithms. Distance from true breakpoint is in bp. Walle (blue) and Pinosio et al. algorithms reach a sensitivity above 0.50 at a distance of 5 bp from true breakpoint, as well as the Merged dataset resulting from their integration (green).

In conclusion, Walle outperforms other tools in term of sensitivity and accuracy both in recall and breakpoint definition. Unlike its competitors, Walle doesn't need a database of annotated transposable elements in order to run, as only the BAM file is sufficient to perform the analysis. However, that can also be a limitation in highly repeated and fragmentary reference assemblies due to issues in the resulting alignment. While other algorithms are also subject to same issues, they are advantaged by the use of a curated TE database. On the other hand, Walle is able to find insertions of not annotated TE and implements a novel detection method based on split reads. An integrated approach could help handling more complex and repeated genomes – as maize is – where detection limits due to misalignments can be compensated by the synergistic application of two or more detection algorithms, which relies on different methods.

4.3 Identification of deletions in *Zea mays*

Deletions compared to the B73 *Zea mays* reference sequence were detected with a combination of Delly (Rausch et al., 2012) and GASV (Sindi et al., 2009), selected based on results obtained in simulations performed in *P. trichocarpa* (Pinosio et al., 2016) and in *V. vinifera* (Gabriele Magris, PhD thesis). As *Zea mays* lines analyzed are mostly homozygous, heterozygous calls were retained only in regions of residual heterozygosity, identified as regions of 100kb with a SNP number five times the median of heterozygous SNPs count in W153R, which is the most heterozygous line.

A total of 48,904 deletions were identified across 6 maize lines of the MM founders, as shown in Table 3.

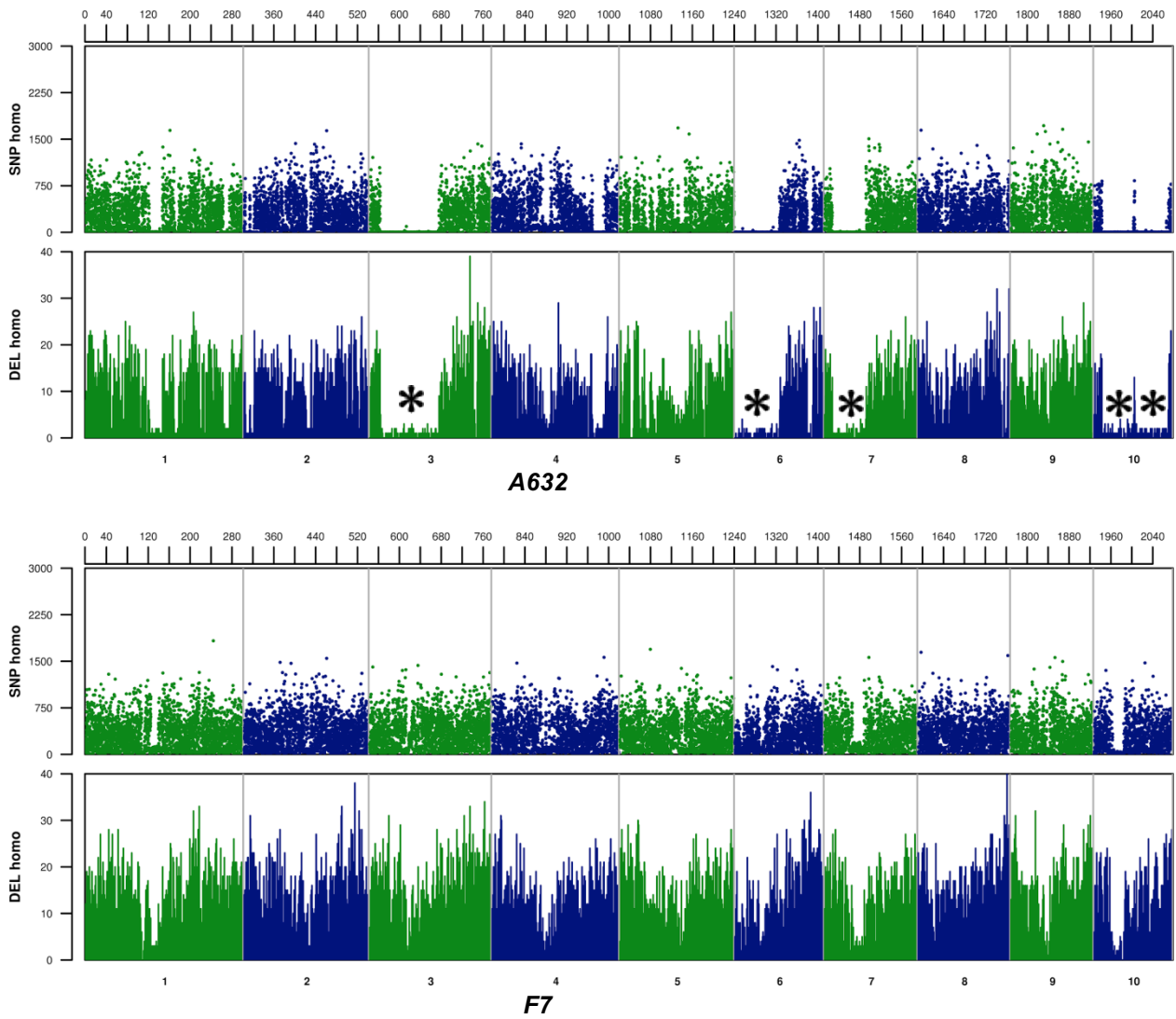
Table 3: Summary of the deletions identified for each variety

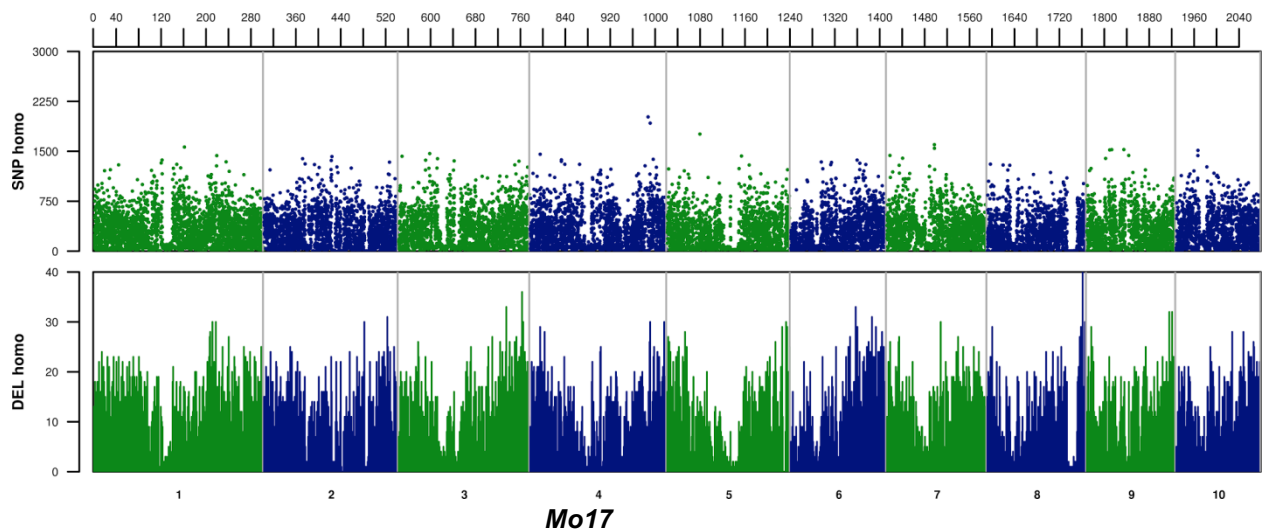
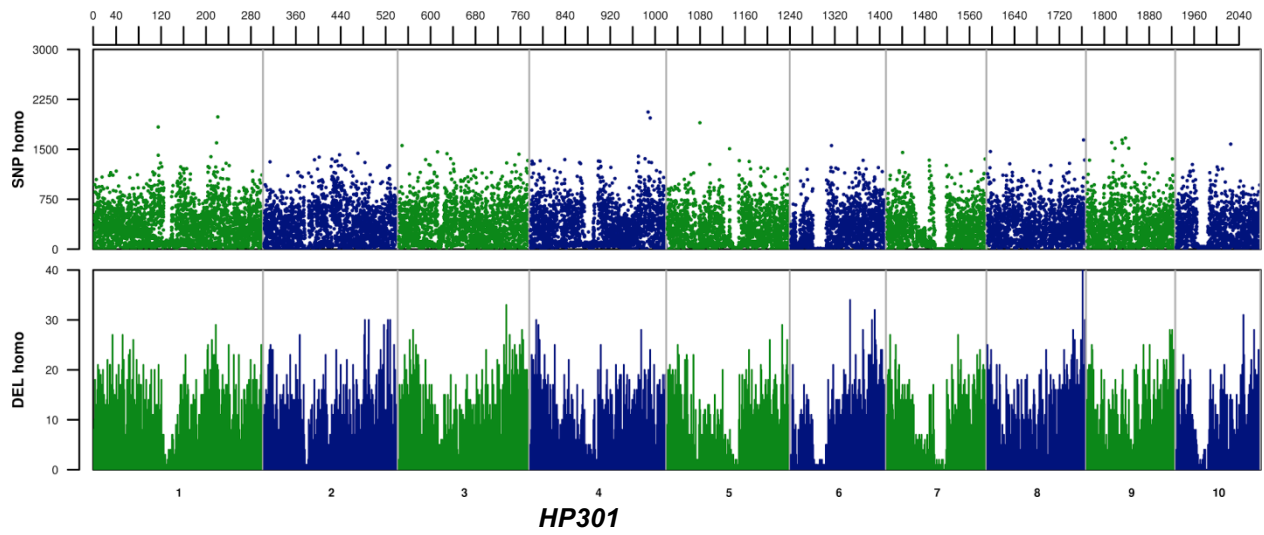
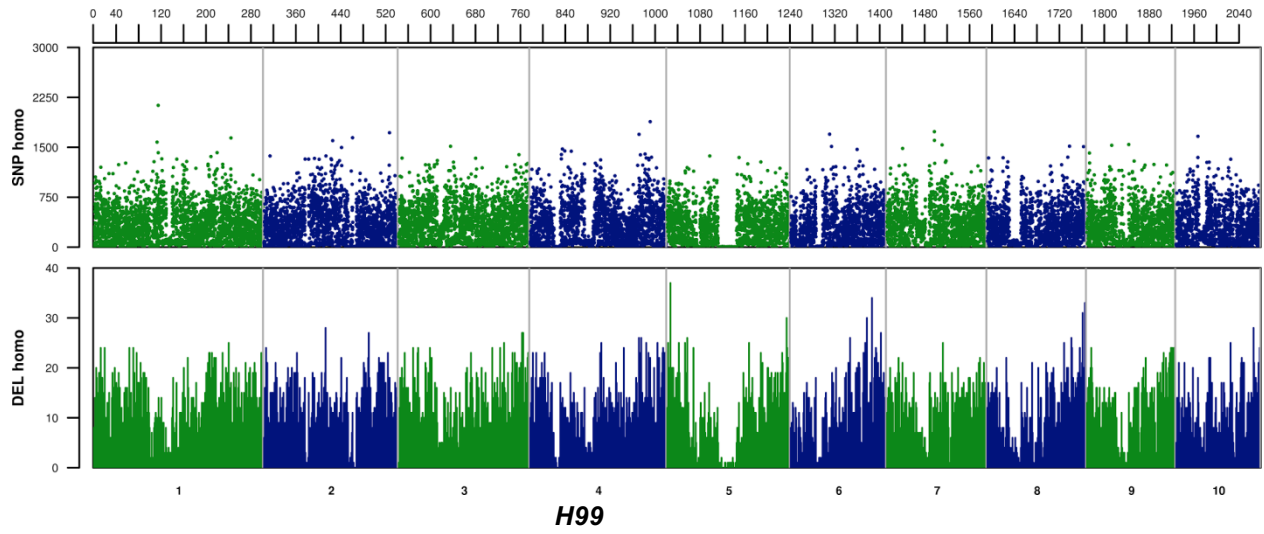
	A632	F7	H99	HP301	Mo17	W153R
Total deletions	17264	26784	21361	23169	23799	21104
Private deletions	1552	4197	1805	2622	2949	1598
Count hom.	17217	26774	21355	23160	23793	20164
Count het.	47	10	6	9	6	940
Length hom. (kb)	195,303,976	305,196,632	242,500,463	261,813,488	272,658,311	227,394,013
Length het. (kb)	531,810	128,104	70,807	78,284	52,876	10,220,151

A632 was the line with the lowest number of both total and private deletions (17,264 and 1,552 respectively), and was the only line in which the total length of deletions was lower than 200 Mb. This was expected as A632 is the most similar line to the B73 reference genotype (Dell'Acqua et al., 2015). F7 was the line with the highest number of both total and private insertions, and the only one in which total length of deletions exceeded the size of 300 Mb.

As W153R is the only line with a considerable level of residual heterozygosity, it was also the line carrying the higher number of heterozygous deletions (940, reaching a size of 10 Mb).

Homozygous deletions and SNPs distributions on chromosomes were plotted for each *Zea mays* line in Figure 17, while heterozygous deletions and SNPs were plotted only for W153R. Windows of 1Mb were used for deletions, while windows of 100 kb were used for SNPs.





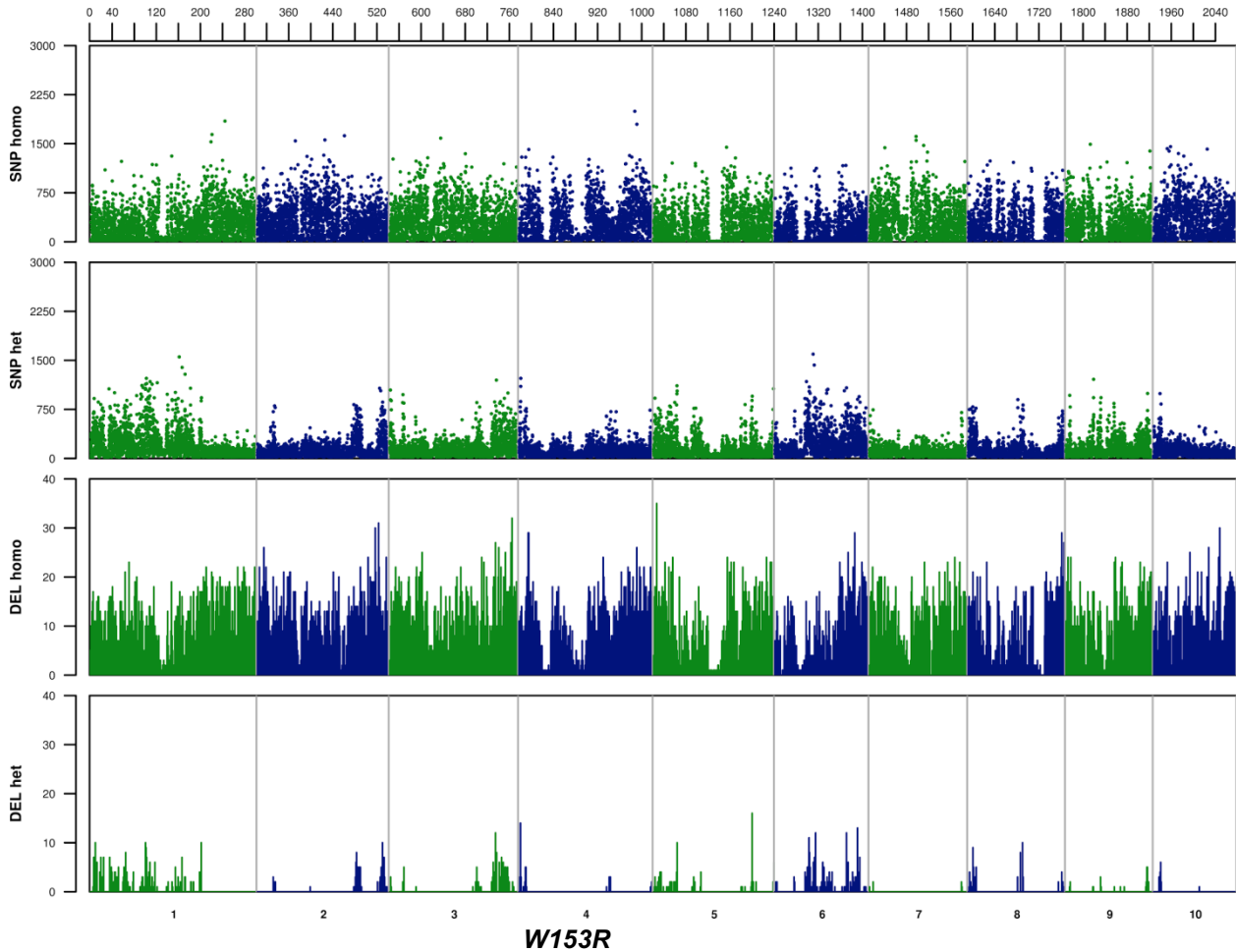


Figure 17. Homozygous deletions and SNPs distributions for each *Zea mays* line plotted across chromosomes from 1 to 10. Heterozygous deletions and SNPs distributions plotted for W153R line only. Large (>50 Mbp) IBD regions in A632 line are marked with an asterisk.

All lines show a large amount of structural and sequence diversity relative to the reference. However, A632 shows large regions of shared IBD with the B73, evidenced by the absence of deletions and SNPs. This is in agreement with the observation that A632 is – among the lines included in this study – the one with the closest genetic relatedness to B73 (Dell’Acqua et al., 2015), though we couldn’t confirm the larger SV reported in the same study for W153R and Mo17 in the short arm of chromosome 6, as our detection is limited to SVs of less than 50 Kb.

A region of ~5 Mb (from 203 Mb to 208 Mb) near the end of the chromosome 1 has a high number of sequences present in the reference and detected as absent in all the other lines, with a total of 154. Another region of 2 Mb (from 210 Mb to 212 Mb) was found not so far in the same chromosome, with approximately 150 Kb of B73 sequences absent in all other lines, and a total of 66 deletions common to at least two lines. Smaller clusters of extant variation on chromosome 1 are located in the very end (298 Mb to 299 Mb), with a total of 29 non-private deletions, and at the opposite side of the chromosome, around 64 Mb (63 Mb to 65 Mb), with a total of 55 non-private deletions detected along 2 Mb.

Other relevant clusters of deletions were identified on:

- chromosome 2 from 200 Mb to 221 Mb, with 483 events and 27.8% of the sequence non-shared with the reference with at least 2 other lines.
- chromosome 3 from 33 Mb to 35 Mb, with more than 1 Mb (at least half of the region) non-shared by the other 5 lines with both B73 and A632 (which is IBD with the reference in that region).
- chromosome 5 from 212 Mb to 214 Mb, with 70 non-private deletions
- chromosome 7 from 163 Mb to 166 Mb, with 67 non-private deletions

Of 48,904 deletions, 7,433 involve genes, and 5,275 involve exons.

A total of 5,614 genes have deleted exons, and a Gene Ontology (GO) analysis was performed on them and reported in Table 4, where top 5 GO terms for each GO category were reported and in Figure 18 the ratio between significant and annotated genes was plotted.

Table 4: Over-represented GO categories influenced by deletions

Category ¹	GO ID ²	Term ³	Signif. ⁴	Annot. ⁵	Ratio ⁶	P value ⁷
BP	GO:0008152	metabolic process	261	2161	0.121	1.10E-08
BP	GO:0044238	primary metabolic process	191	1494	0.128	2.50E-08
BP	GO:0019538	protein metabolic process	75	451	0.166	3.90E-08
BP	GO:0071704	organic substance metabolic process	199	1596	0.125	8.30E-08
BP	GO:0043170	macromolecule metabolic process	133	977	0.136	1.30E-07
CC	GO:0044425	membrane part	59	418	0.141	5.80E-06
CC	GO:0016021	integral component of membrane	44	280	0.157	5.90E-06
CC	GO:0031224	intrinsic component of membrane	44	285	0.154	9.40E-06
CC	GO:0016020	membrane	123	1168	0.105	3.00E-04
CC	GO:0005634	nucleus	49	398	0.123	9.40E-04
MF	GO:0003674	molecular_function	1887	16666	0.113	2.50E-13
MF	GO:0003824	catalytic activity	682	5373	0.127	7.90E-13
MF	GO:0003676	nucleic acid binding	357	2846	0.125	8.40E-07
MF	GO:0016772	transferase activity, transferring phosphorus-containing groups	164	1182	0.139	4.20E-06
MF	GO:0016740	transferase activity	246	1908	0.129	6.20E-06

¹ Gene ontology category domain: BP, Biological Process; CC, Cellular Component; MF, Molecular Function.

² Gene ontology term ID. ³ Gene ontology term name. ⁴ Significant annotated genes influenced by deletions found enriched for each GO term. ⁵ Annotated genes for each GO term. ⁶ Ratio between significant and annotated genes for each over-represented gene ontology category. ⁷ Fisher's exact test p value of the enrichment of GO terms with genes affected by deletions. All GO terms reported were significant ($p < 0.01$).

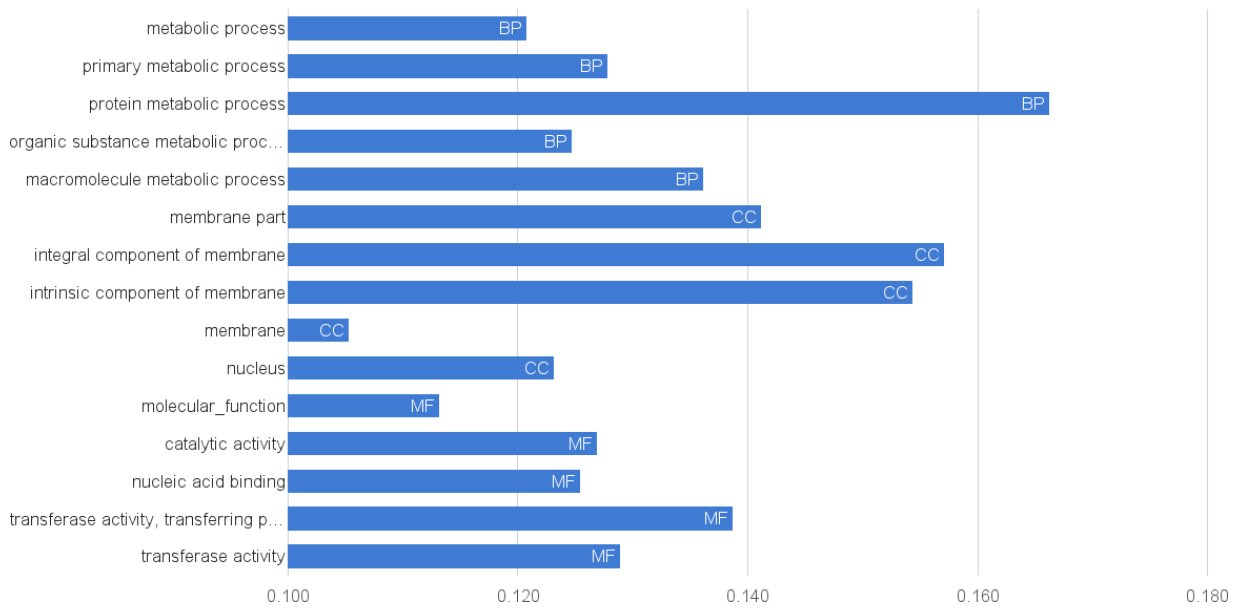


Figure 18. Significant processes enriched in genes influenced by deletions. On the left the top 5 terms per each GO category (BP, Biological Process; CC, Cellular Component; MF, Molecular Function). Bars length represent the ratio between the observed significant genes and all annotated genes.

The most influenced GO terms are *protein metabolic process* term, accounting for 16% of the annotated genes interrupted by deletions, and component of membrane terms (same genes enriched in 2 redundant terms, plus a third more general *membrane* term found), accounting for 14-15 % of the annotated genes interrupted by deletions. The first gene category identified is reflected by the tendency of TEs to be associated to genes that encode important enzymes for transposition and integration (Gao et al., 2012) and may correspond to TE element erroneously annotated as genes.

A comparison of GO results with another analysis made on transcript assemblies of 503 maize inbred lines (Hirsch et al., 2014) reveals that some of the most frequently observed terms are common in both analyses - in particular, membrane associated, nucleotide binding, and catalytic activity terms.

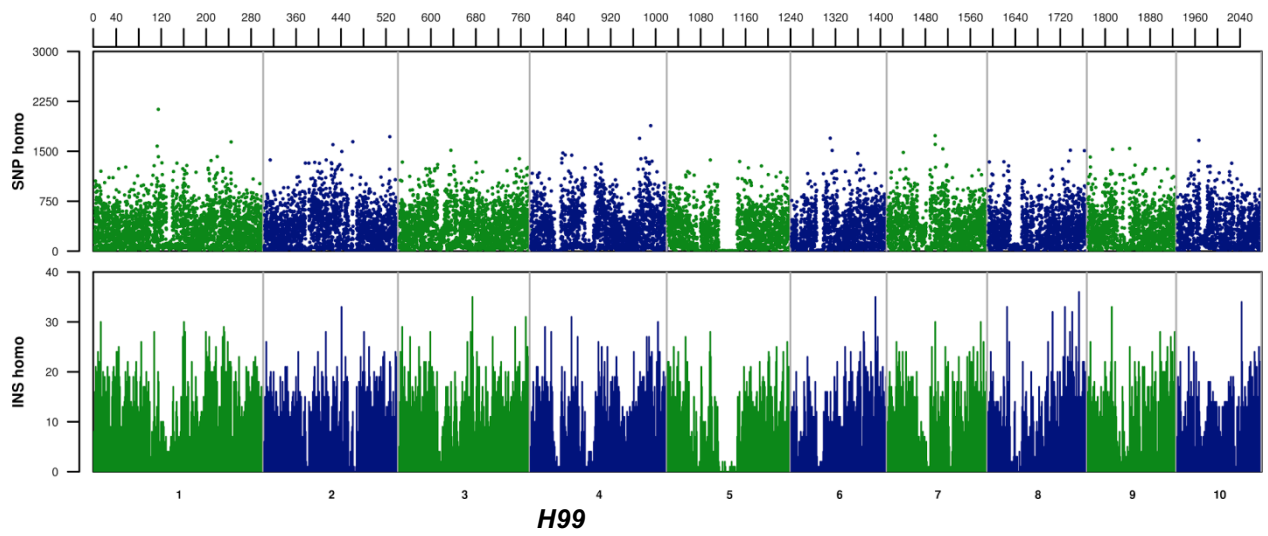
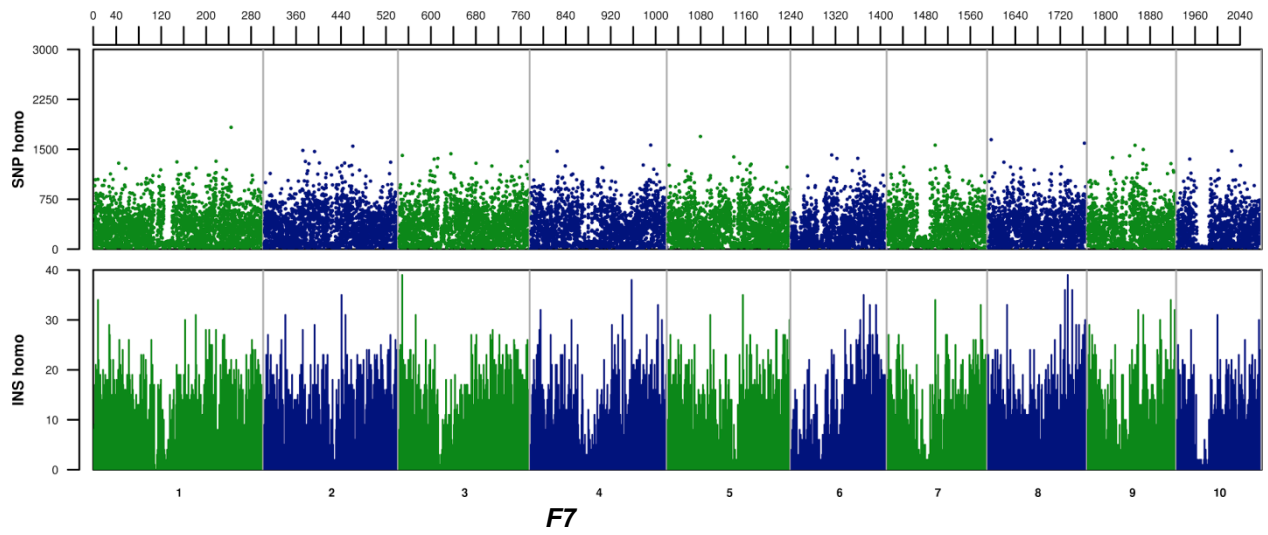
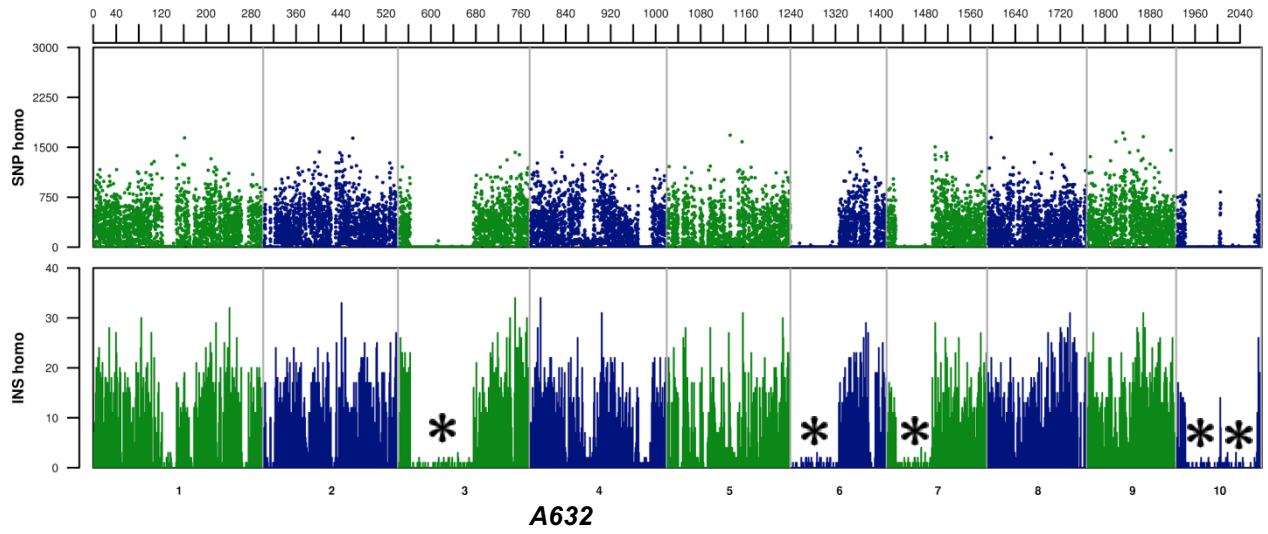
4.4 Identification of insertions in *Zea mays*

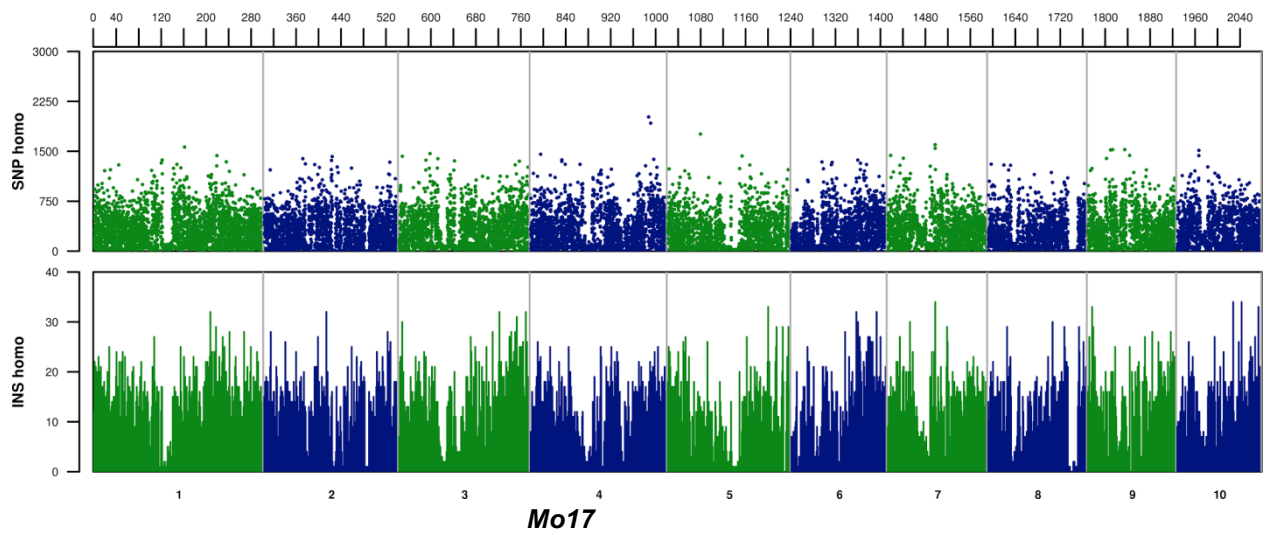
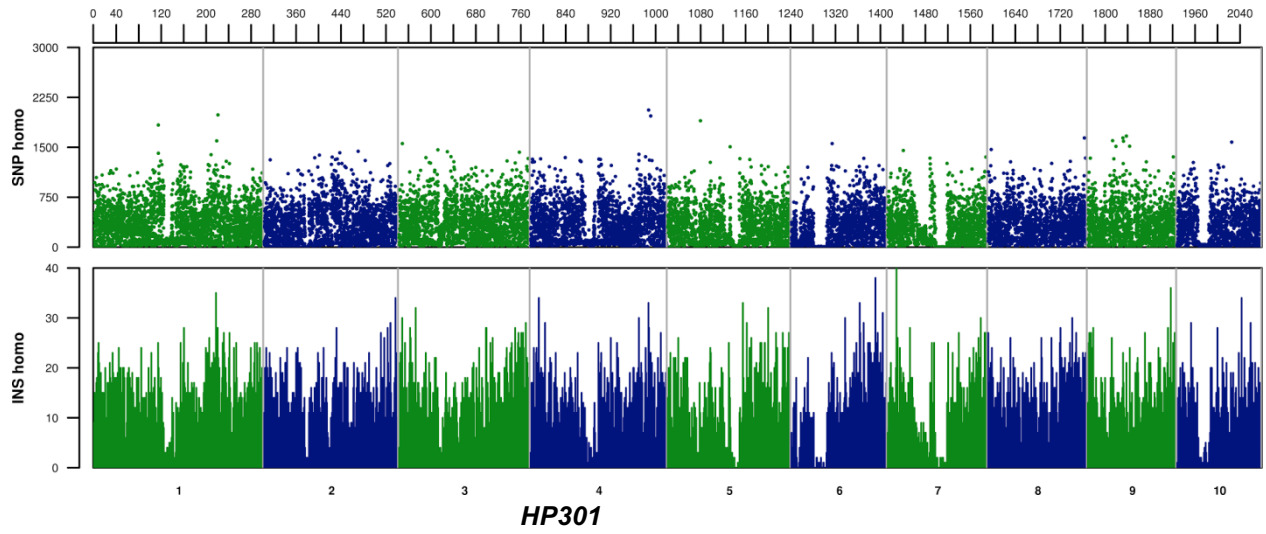
The Merged approach was used as previously described to detect insertions compared to the B73 *Zea mays* reference sequence. The same heterozygous filter as for the deletions was applied, retaining heterozygous calls only on regions of residual heterozygosity. A total of 75,370 insertions were identified across 6 maize lines of the MM founders, as shown in Table 5.

Table 5: Summary of the insertions identified for each variety

	A632	F7	H99	HP301	Mo17	W153R
Total insertions	18887	28142	24387	25319	24741	22877
Private insertions	4628	9825	5394	7287	7001	4216
Count hom.	18819	28115	24351	25307	24720	21578
Count het.	68	27	36	11	21	1299

In line with deletions, A632 was the line with the lowest number of both total insertions, though W153R has less private insertions. Again, F7 was the line with the highest number of both total and private insertions. Homozygous insertions and SNPs distributions on chromosomes were plotted for each *Zea mays* line in Figure 19, while heterozygous insertions and SNPs were plotted only for W153R. Windows of 1Mb were used for insertions, while windows of 100 kb were used for SNP.





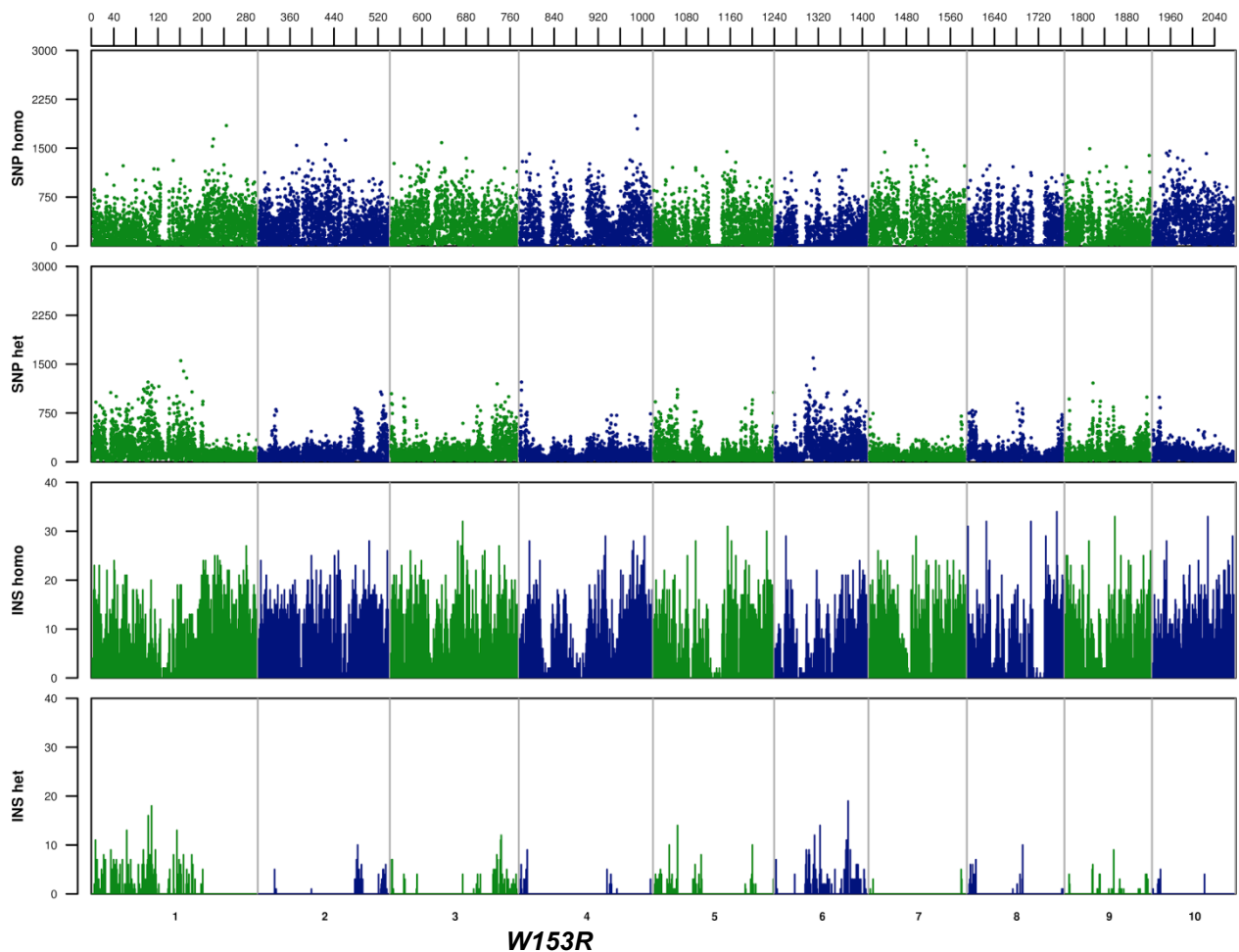


Figure 19. Homozygous insertions and SNPs distributions for each *Zea mays* line plotted across chromosomes from 1 to 10. Heterozygous insertions and SNPs distributions plotted for W153R line only. Large (>50 Mbp) IBD regions in A632 line are marked with an asterisk.

Two major clusters of insertions were detected for chromosome 1, from 199 Mb to 208 Mb and from 231 Mb to 236 Mb, respectively with 297 and 152 non-private insertions. This is in line with what was observed for deletions, where the regions from 203 Mb to 208 Mb showed a high density of deletions. Chromosome 3 presents clusters of insertions at both its extremities (from 2 Mb to 6 Mb, and from 206 Mb to 210 Mb). Also chromosome 5 presents clusters of insertions (from 204 to 206 Mb, with 83 insertion events common to at least two lines) near the respective region as for deletions.

Taken together, results from both insertions and deletions allowed us to identify some high-variability regions, mostly concentrated on some chromosome than others, in particular at the end of chromosome 1 – especially in the 2 Mb from 206 Mb to 208 Mb, where both type of SVs were detected - and at the end of chromosome 5 (from 206 to 214 Mb).

Of 75,370 insertions, 8,775 involved genes, suggesting that insertions occur mostly in intergenic portions. 2,488 insertions interrupt exons, which is slightly more than 3% of all detected insertions. A total of 2,951 genes have exons interrupted by an insertion, and a Gene Ontology (GO) analysis was performed on them and reported in Table 6, where top 5 GO terms for each GO category were reported. In Figure 20 the ratio between significant and annotated genes was plotted.

Table 6: Over-represented GO categories influenced by insertions

Category ¹	GO ID ²	Term ³	Signif. ⁴	Annot. ⁵	Ratio ⁶	P value ⁷
BP	GO:0000377**	RNA splicing, via transesterification reactions ...	14	57	0.246	1.10E-03
BP	GO:0000375**	RNA splicing, via transesterification reactions	14	58	0.241	1.30E-03
BP	GO:0000398**	mRNA splicing, via spliceosome	11	46	0.239	4.50E-03
BP	GO:0000255**	allantoin metabolic process	2	2	1.000	9.80E-03
BP	GO:0000256**	allantoin catabolic process	2	2	1.000	9.80E-03
CC	GO:0031968*	organelle outer membrane	3	8	0.375	2.20E-02
CC	GO:0005773*	vacuole	22	173	0.127	2.30E-02
CC	GO:0005777*	peroxisome	7	40	0.175	3.90E-02
CC	GO:0042579*	microbody	7	40	0.175	3.90E-02
CC	GO:0031300	intrinsic component of organelle membrane	3	11	0.273	5.30E-02
MF	GO:0004747**	ribokinase activity	4	6	0.667	6.60E-04
MF	GO:0003824**	catalytic activity	516	5373	0.096	8.50E-04
MF	GO:0019200**	carbohydrate kinase activity	10	38	0.263	9.30E-04
MF	GO:0016491**	oxidoreductase activity	97	849	0.114	1.46E-03
MF	GO:0016209**	antioxidant activity	31	213	0.146	2.02E-03

¹ Gene ontology category domain: BP, Biological Process; CC, Cellular Component; MF, Molecular Function.

² Gene ontology term ID (* $p < 0.05$; ** $p < 0.01$). ³ Gene ontology term name. ⁴ Significant annotated genes influenced by insertions found enriched for each GO term. ⁵ Annotated genes for each GO term. ⁶ Ratio between significant and annotated genes for each over-represented gene ontology category. ⁷ Fisher's exact test p value of the enrichment of GO terms with genes affected by insertions.

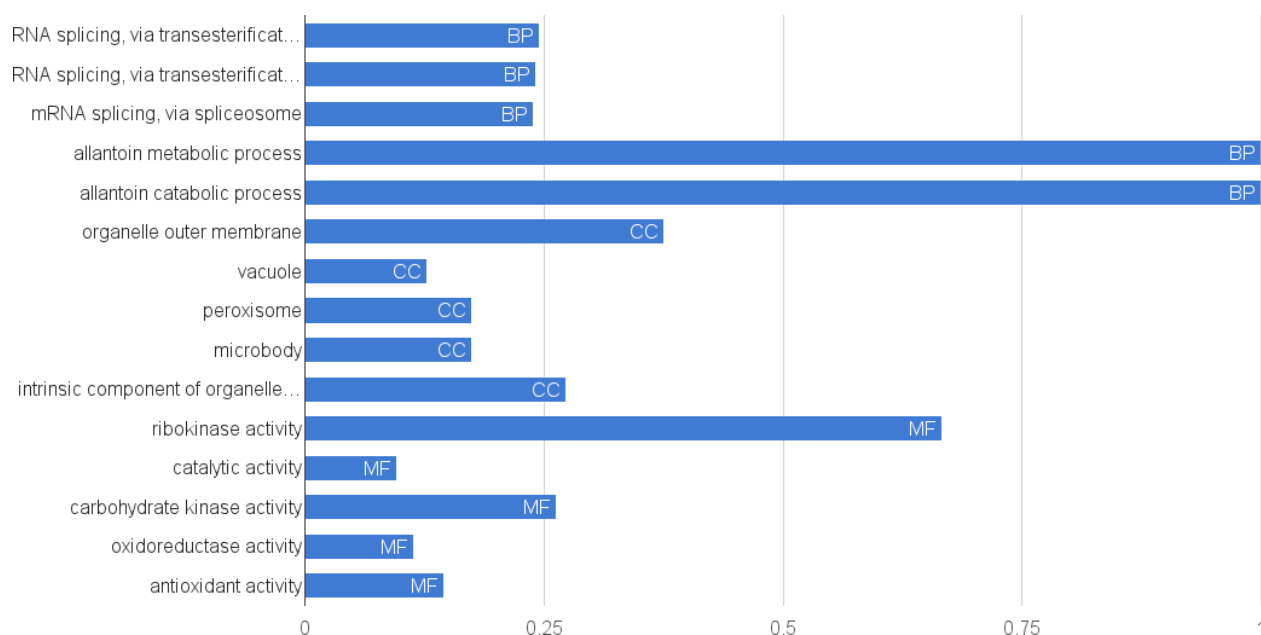


Figure 20. Significant processes enriched in genes influenced by insertions. On the left the top 5 terms per each GO category (BP, Biological Process; CC, Cellular Component; MF, Molecular Function). Bars length represent the ratio between the observed significant genes and all annotated genes.

All genes involved in allantoin metabolic and catabolic processes were found to be broken by an insertion, hence they are just 2. Nevertheless, they have been associated to antioxidant activity (Nourimand et al., 2016) and a total of 31 genes associated to such term (14% of the annotated genes) are found broken by an insertion, too. Moreover, *ribokinase activity* and *carbohydrate kinase activity* terms accounting respectively 66% and 25% of the annotated genes interrupted by insertions and similar terms were also found to be present in representative transcript assemblies of another study (Hirsch et al., 2014).

4.5 Classification of SV

In order to characterize the dispensable genome composition, an analysis of sequences of deletions was performed in terms of: the overall composition of TEs, and the fraction of deletions annotated as complete elements.

The first analysis aimed at investigating an overall composition of the dispensable genome, while the second allows discriminating real deletions from insertions in the reference, as deletions annotated as complete elements are probably insertions in the reference in consideration of the predominant copy and paste mode of transposition of TEs in plants.

The deletions length distribution (Figure 21) gives an idea of which type of elements are present.

Four peaks could be identified at 1-1.5 kb, 7-7.5 kb, 9-9.5 kb, and 13.5-14 kb. The first peak represents a large fraction of deletions composed by small class II TEs or portions of class I TEs; the second peak corresponds to the average size of LTR class I TEs belonging to Copia (RLC) superfamily; the third peak corresponds to the average size of LTR TEs belonging to Gypsy (RLG) superfamily; the fourth peak represents nested elements composed by RLC superfamily but not RLG superfamily, and by a more detailed nested elements analysis (see 4.6) it results that such elements have an average length of 14 kb.

Interestingly, the size of the fourth peak is approximately the double of the peak associated to RLC superfamily, suggesting some sort of preference in nesting structures formation, and this was further investigated. Moreover, a small group of deletions can be observed at 22.5-23 kb, which is in line with the average length observed for nested elements (see 4.6).

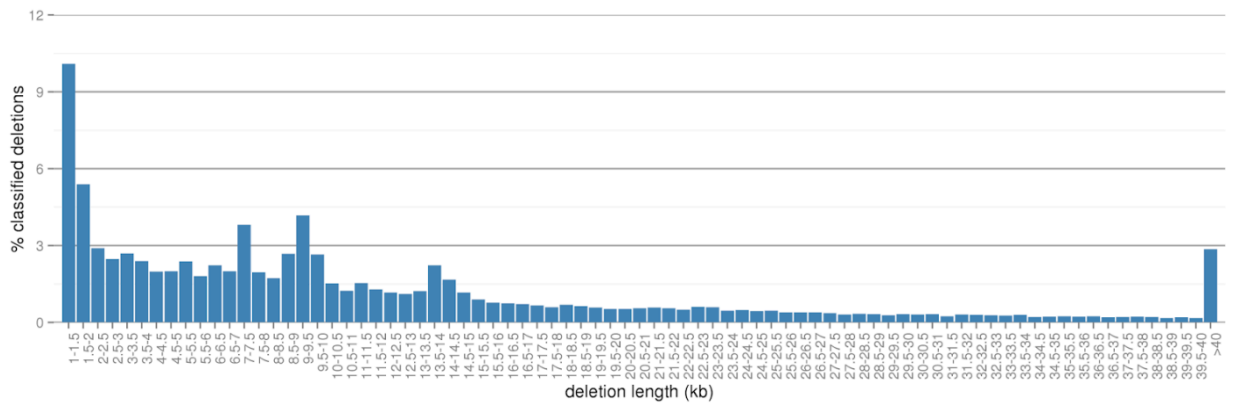


Figure 21. Length distribution of deletions with bins of 0.5 kb (calculated as percentage of all deletions length). Since our detection is focused on deletions larger than 1 kb, the first bin is from 1 to 1.5 kb, and represents small class II TEs or portions of class I TEs. Other peaks are at 7-7.5 kb (LTR Copia elements), 9-9.5 kb (LTR Gypsy elements), 13.5-14 kb (nested elements composed by RLC superfamily), and 22.5-23 kb (other nested elements).

All the deletions were characterized at first by their TE compositions in bp, using a curated TE database of *Zea mays* (Wessler et al., <http://maizetadb.org/>) of 1526 non-redundant elements, dated 12th February 2015.

Our results (Table 7, Figure 22) highlight how the database lacks in Helitron elements (only 16 sequences present as DHH), and consequently our annotation underestimates such elements.

In fact, by masking the genome with the TE database, we were able to find less than 1% of Helitrons genome-wide, while it was estimated that Helitrons composed from 2% to 6% of the maize genome (Yang and Bennetzen, 2009; Xiong et al., 2014). However, considering the public annotation of *Z. mays* (RefGen_v3 annotation build 5b+), a similar fraction of Helitrons in the genome is reported.

Most of the deletions (80%) were due to LTR-retrotransposons, suggesting their active role in maintaining the maize genome in a flux (Brunner et al., 2005; Morgante et al., 2005).

Table 6: Deletions composition in TE superfamilies

<i>Superfamily</i>	<i>bp</i>	<i>% total deletions</i>	<i>% genome-wide TE abundance</i>
<i>DHH</i>	5,081,121	0.90%	0.44%
<i>DTA</i>	6,379,294	1.13%	1.22%
<i>DTC</i>	15,328,684	2.72%	2.85%
<i>DTH</i>	4,561,640	0.81%	0.82%
<i>DTM</i>	6,079,768	1.08%	0.93%
<i>DTT</i>	305,746	0.05%	0.15%
<i>RIL</i>	6,395,847	1.13%	0.61%
<i>RIT</i>	379,404	0.07%	0.09%
<i>RLC</i>	167,442,143	29.68%	24.25%
<i>RLG</i>	269,889,623	47.84%	45.32%
<i>RLX</i>	14,954,044	2.65%	4.63%
<i>RST</i>	64,389	0.01%	0.05%
<i>not_masked</i>	67,259,420	11.92%	-

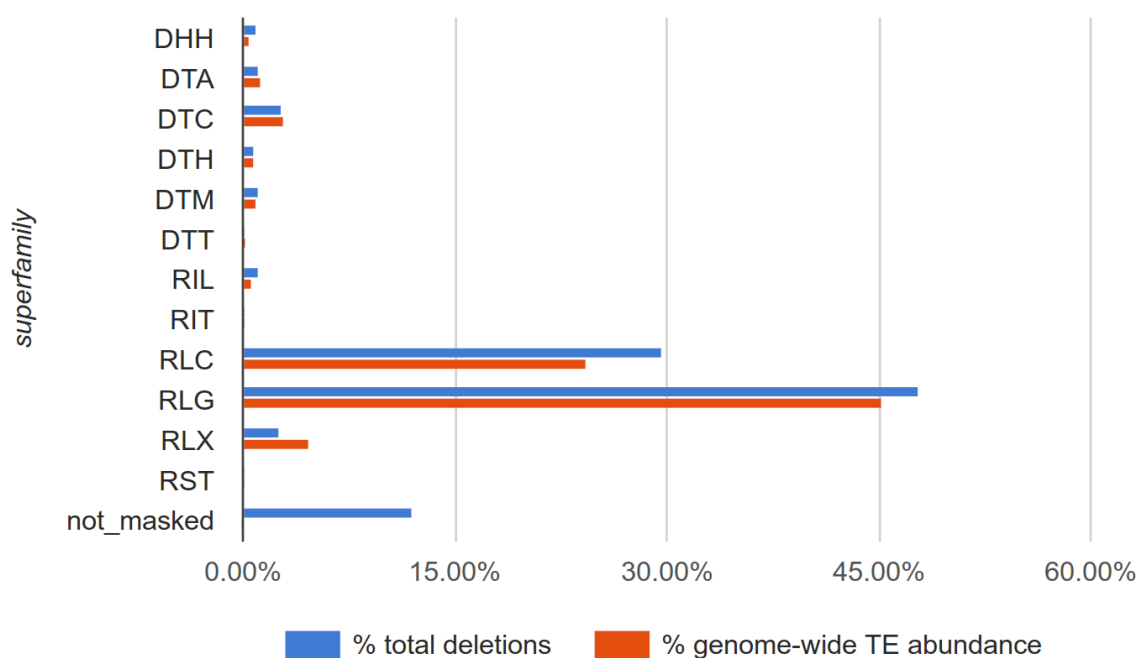


Figure 22. Deletions composition is shown in term of annotated sequences as TE superfamilies (blue bars), in comparison to same TE superfamily genome-wide abundance (red bars). *not_masked* category includes deletion sequences not identified as a TE element.

RLG superfamily (Gypsy) is the most abundant superfamily both in deletions and genome-wide, followed by RLC superfamily (Copia). LTR elements are the 80% of deletions and about the 75% of TE in the genome.

LTR retrotransposons were further investigated at the family level, to identify the most active families (Table 7, Figure 23).

For RLX superfamilies, *milt* seems more mobile than *ruda* family, which is more present genome-wide, but in deletions the ratio is 2:1. Along the RLG and RLC superfamilies, larger families are also the most mobile, which justifies their spread in the genome.

Table 7: Deletions composition in LTR families

LTR superfamily	LTR family	bp	% total deletions	% genome-wide TE abundance
RLX	<i>milt</i>	7,882,565	1.40%	1.11%
RLX	<i>ruda</i>	3,702,827	0.66%	1.13%
RLG	<i>huck</i>	87,116,510	15.44%	11.34%
RLG	<i>cinful-zeon</i>	53,239,546	9.44%	9.12%
RLG	<i>prem1</i>	22,477,387	3.98%	3.84%
RLG	<i>xilon-diguus</i>	20,308,228	3.60%	4.25%
RLG	<i>grande</i>	18,225,056	3.23%	3.12%
RLG	<i>flip</i>	15,797,409	2.80%	4.78%
RLG	<i>gyrna</i>	11,657,043	2.07%	3.27%
RLG	<i>doke</i>	10,550,022	1.87%	1.87%
RLG	<i>puck</i>	5,259,451	0.93%	1.11%
RLG	<i>uwum</i>	4,151,845	0.74%	0.79%
RLG	<i>tekay</i>	3,376,296	0.60%	0.80%
RLG	<i>dagaf</i>	3,257,119	0.58%	0.80%
RLC	<i>ji</i>	86,123,701	15.27%	11.09%
RLC	<i>opie</i>	53,260,433	9.44%	9.22%
RLC	<i>giepum</i>	7,223,715	1.28%	1.59%
RLC	<i>gudyeg</i>	2,057,405	0.36%	0.21%
RLC	<i>wiwa</i>	1,971,703	0.35%	0.34%
RLC	<i>machiavelli</i>	1,727,407	0.31%	0.18%
RLC	<i>ebel</i>	1,323,633	0.23%	0.22%
RLC	<i>eninu</i>	1,087,704	0.19%	0.09%
RLC	<i>raider</i>	703,244	0.12%	0.10%
RLC	<i>fourf</i>	571,852	0.10%	0.23%
RLC	<i>doke</i>	549,445	0.10%	0.50%
RLC	<i>debeh</i>	513,779	0.09%	0.05%
RLC	<i>japov</i>	501,005	0.09%	0.02%

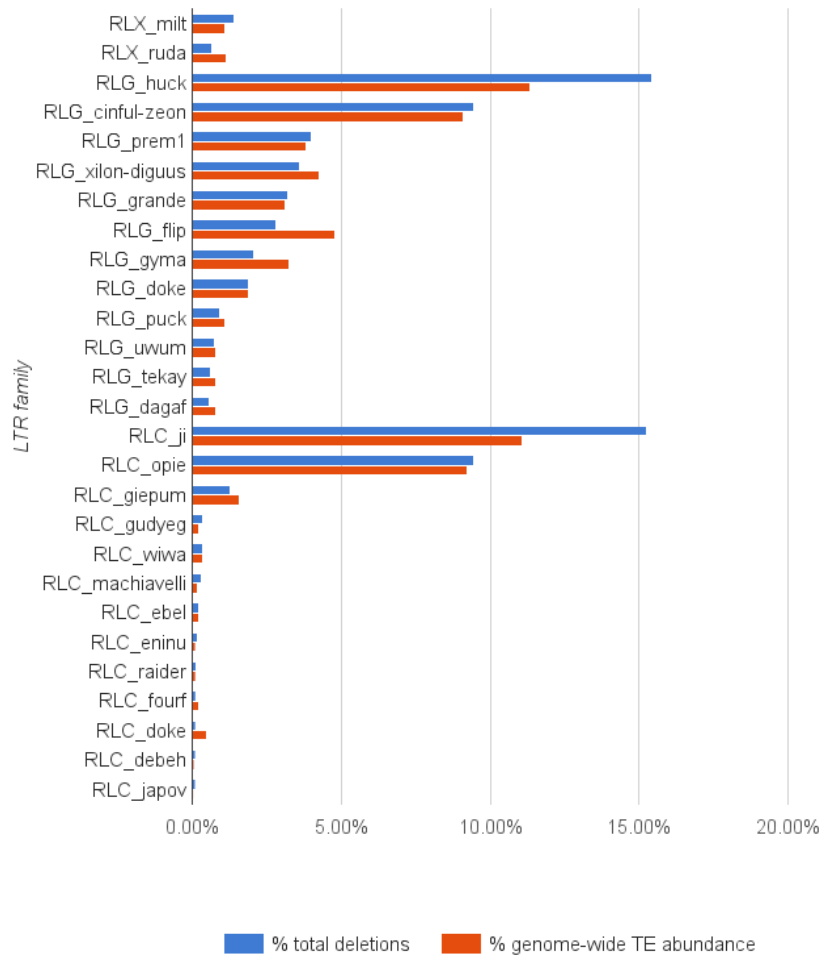


Figure 23. Deletions composition is shown in term of annotated sequences as LTR families (blue bars), in comparison to the genome-wide TE abundance of the same LTR family (red bars).

RLG_huck and *RLC_ji* are the most abundant LTR elements, followed by *RLC_cinful-zeon* and *RLC_opie*. Together, they are the 50% of deletions annotated as TE, while only about 40% of TE in the genome.

Our internally developed pipeline for the annotation developed by Michele Vidotto and Gabriele Magris was used to annotate deletions.

Annotations may result in complete elements, i.e. deletions in which both edges corresponded to edges of TEs of the same superfamily or having an 80% match with a specific transposable element, or in incomplete elements (i.e. non-classified deletions, without any matches or matching only partly with a specific TE superfamily).

We have identified 30,513 complete elements, for approximately 357 Mb of sequence (63.28 % of deletions), while the remaining 18,391 (summing up to 207 Mb of sequence) consists in incomplete elements.

Of all the complete elements, 9,735 are nested (31.9% of complete elements), and 9.2% contain gene fragments (of at least 50bp), accounting for less than 3Mb of sequence. Incomplete elements present complex nesting patterns and 31% contain gene fragments for a total size of 7.9 Mb.

Incomplete elements might include elements missing from the annotation (i.e. Helitrons) and actual deletions, while most of complete elements are probably insertions in the reference haplotype, since the definition of deletion has only a technical meaning to indicate a sequence present in the reference and absent in at least another maize line. This consideration is supported by the fact that the greatest fraction of complete elements identified mobilize themselves through a copy-and-paste transposition mechanism.

Superfamilies classification of complete elements is shown in Table 8. Only the count is shown, as the full length of the element doesn't fit with the length of the deletion, in case of nested events.

Table 8: Classification of complete elements

Superfam.	count	% of complete deletions
DHH	1,176	3.85%
DTA	1,648	5.40%
DTC	1,778	5.83%
DTH	967	3.17%
DTM	1,021	3.35%
RIL	923	3.02%
RLC	8,073	26.46%
RLG	12,815	42.00%
RLX	2,025	6.64%
others	87	0.28%

The fraction of class I elements has increased relative to their overall fraction, suggesting they are present mostly as complete and active elements. LTR Gypsy (RLG) are the most abundant complete element, followed by LTR Copia superfamily (RLC). Although Helitron superfamily (DHH) is in proportion more abundant as complete elements than they are without stratifying by completeness, they are still underestimated.

In order to annotate insertions, we used both the full set of deletions and the downloaded maize TE database to perform our analysis.

Since a detailed study of the inserted sequence is not possible as it is an inferred estimation, only the count of annotated elements is reported in Table 9.

The unknown fraction includes not annotated portion of incomplete deletions and Walle calls not annotated as a complete TE.

Table 9: Classification of insertions

Superfam.	count	% of insertions
DHH	1422	1.89%
DTA	4414	5.86%
DTC	4936	6.55%
DTH	3351	4.45%
DTM	891	1.18%
RIL	1429	1.90%
RLC	17291	22.94%
RLG	28619	37.97%
RLX	6505	8.63%
others	723	0.96%
unknown	5789	7.68%

The timing of insertion of elements belonging to LTR-retrotransposon superfamily was estimated by comparing their LTR sequences. At the time when a LTR retrotransposon transposes itself, its two LTR sequences are identical, and they accumulate mutations over time.

In Figure 24 is shown the comparison between the time of insertion for annotated LTR elements detected as deletions (polymorphic LTRs), and the time of insertions of LTR elements annotated in the reference genome and not participating in SV (non-polymorphic LTRs). While there is a detectable and significant difference between the distributions of insertion times of the two sets of elements, the difference is less remarkable than that observed by Brunner *et al.* (2005) on the basis of the Sanger sequencing of a set of BAC clones in B73 and Mo17 inbred lines. This may reflect a higher rate of false negatives in our analysis of deletions based on NGS due to the inability to map short reads in repeated sequences and thus to detect deletions of TEs within such repeated regions. False negatives would put polymorphic insertions of younger elements together with non-polymorphic insertions in our analysis. The presence of a high rate of false negatives could also account for a lower total number of deletions detected in comparison to genome-wide expectations based on the above-mentioned Sanger sequencing of a limited number of BAC clones (Brunner *et al.*, 2005).

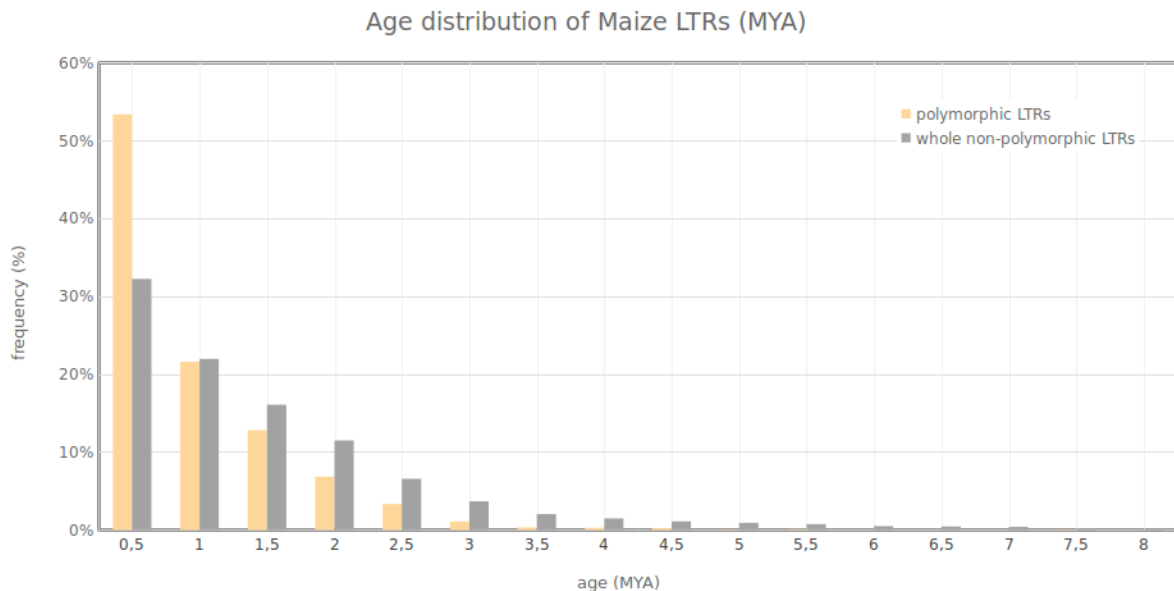


Figure 24. Insertion times (Mya) of LTR retrotransposons involved in structural variation polymorphic LTRs) and of LTR elements not involved in structural variation (non-polymorphic LTRs).

The barplot shows an increasing of LTRs expansion in the last 3 Mya, which is a time consistent with previous evidences (SanMiguel et al., 1998).

It is strikingly shown that while the genome fraction of LTR expands, those elements progressively invade it, forming the complex structures of nesting elements of retrotransposons that we found in the maize mobile fraction of the genome.

4.6 Nested elements analysis

About one third of detected deletions (31.1%) are composed by nested elements so their composition and patterns were further investigated.

Nested elements include a wide range of structures, from relatively simple linear nesting one-into-another TEs (defined as A-B-A structures, or one-level nesting), to more complex structures, with an average length of 22 kb.

Frequencies of each superfamily to participate in nesting structures were evaluated by the Pearson's chi-squared test. The null hypothesis was that the observed frequency distribution is consistent with the distribution of frequencies of all other superfamilies to participate in nesting structures. Such frequencies are represented in Figure 25.

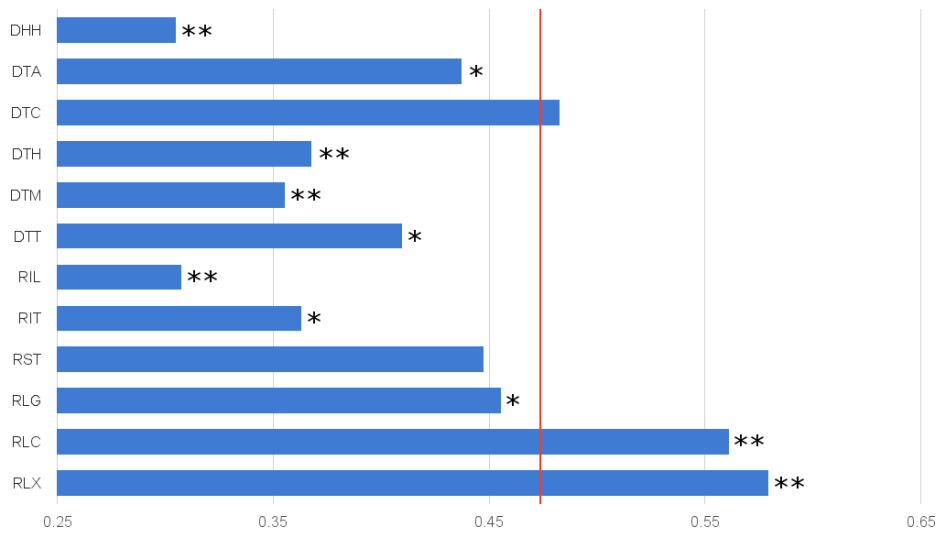


Figure 25. Frequencies of each superfamily to participate in nesting structures.

The red line indicates the frequency of all indistinct TEs to participate in nesting structures; ** most significant frequencies ($p = 2.20E-16$); * other significant frequencies ($p < 0.05$).

RLC superfamily has a higher than expected frequency of nesting, as already anticipated by the observation of deletions lengths distribution. Conversely, Helitrons (DHH) and LINEs (RIL) are involved in a lower proportion of nesting events compared to genome average. Interestingly, Mutator (DTM) elements is the most unlikely nesting category of TIR Class II DNA transposons and a similar behavior of such elements was reported in previous studies (Zhao et al., 2014), in which they found that most of DTM containing a nested insertion in maize are from LTR retrotransposons, and argued that the opposite situation is less likely to happen. That is because LTRs are more abundant in the maize genome and they prefer to replicate themselves inside repetitive regions (Jiang and Wessler, 2001) - which DTM transposons are, while the latter prefers to replicate themselves into genic regions (Liu et al., 2009). This is consistent with the findings that LTR *Copia* elements have a higher than expected frequency of nesting, and such structures are the most abundant amongst the nested ones, with a length of 14 kb (Figure 21).

Of 15,226 nesting structures identified, approximately 36% (5,480) present a one level linear A-B-A structure; 2 inner levels of nesting (A-B-C-B-A structures, one element inside an element which is inserted in turn in a third element) are also present in 9% (1,420) of the total nested elements; 2 serial levels of nesting structures (A-B-A-C-A structures, two elements inside a third host element) represent 7.8% (1,187) of nesting structures, while the remaining 7,139 showed more complex multi-level and diversified nesting structures.

Linear A-B-A nested structures were further investigated in order to study the nesting biases among most abundant TE superfamilies.

A heatmap (Figure 26) reports for each TE superfamilies the number of nested TE found inside a host TE, and vice versa.

For each pair of different superfamilies (generically named here A and B), we tested for asymmetry of nesting patterns against the null hypothesis of observing an equal number of events in which A is the host and B is the nested (A-B-A) and events in which B is the host and A is the nested (B-A-B), assuming a Poisson distribution. Deviations from the null means that one of the two superfamilies has a higher tendency of being nested compared to the other.

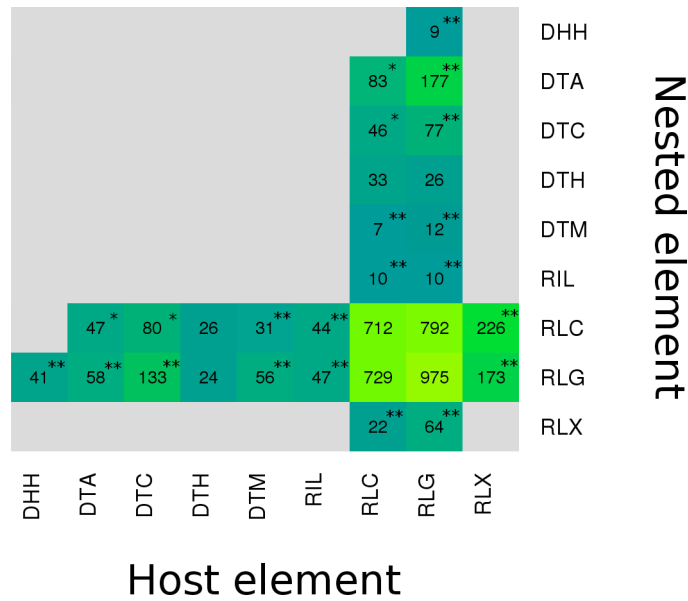


Figure 26. Heatmap of nesting events for each TE superfamily, classified as host or nested element (** $p < 0.01$; * $p < 0.05$). Same superfamily events not tested; only structures annotated for a total of 75% of their sequences were taken into account. Only couples of superfamilies with a sum of events greater than 35 were considered.

Among LTR elements, Gypsy (RLG) and Copia (RLC) superfamilies did not show significant differences from the null, while RLX superfamily tended to act as host for other LTRs ($p < 2E10^{-6}$).

Among TIR Class II DNA elements, hAT elements (DTA) seem to locate inside host LTRs. 260 DTA were nested into LTR (177 in RLG and 83 in RLC), compared to 105 events in which LTR (58 RLG and 47 RLC) were nested into DTA. CACTA (DTC) and Mutator (DTM) elements act as hosts for nested LTRs, in particular for nested Gypsy (respectively 133 and 56 events, $p < 4E10^{-3}$ and $p < 2E10^{-4}$). That is in line with previous evidences showing that DNA transposons - and Mutator superfamilies - are often host for nested LTR elements, though they are unlikely prone to form nesting structures in general, as discussed above (Zhao et al., 2014).

Given the abundance of LTR families, we repeated the analysis stratifying LTR into families. We show results for the 6 families that are most often involved in nesting events. Figure 27 shows for each LTR family the number of nested TE found inside a host TE, and vice versa. We tested for asymmetry of nesting patterns against the null hypothesis of observing an equal number of events for each pair of different LTR families - assuming a Poisson distribution - as described for superfamilies.

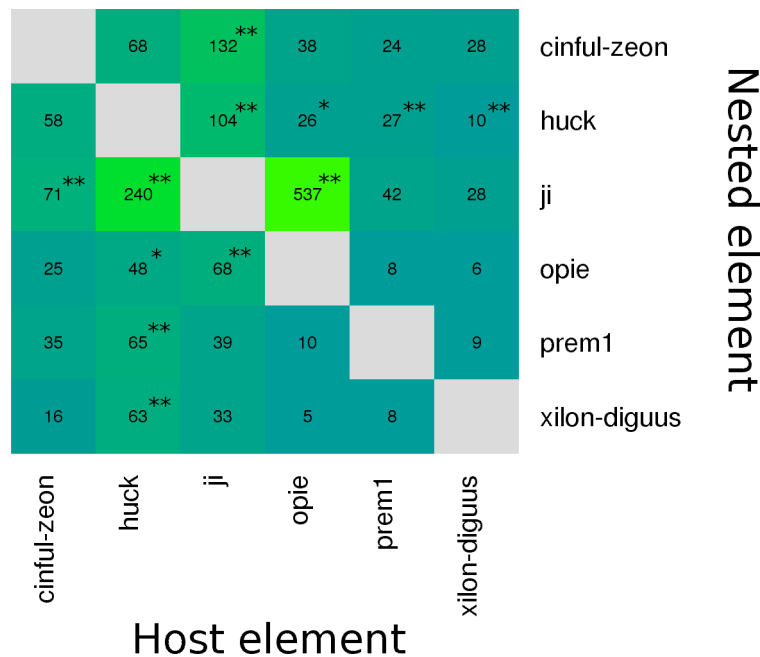


Figure 27. Heatmap of nesting events for most nesting LTR family, classified as host or nested element (** $p < 0.01$; * $p < 0.05$).). Same family events not tested; only structures annotated for a total of 75% of their sequences were taken into account.

In line with previous analysis (Kronmiller and Wise, 2008), ji opie and huck are the most involved in nested structures and RLC opie are likely to contain most of other RLC ji (537 events, $p < 1E-23$), together with RLG huck (240 events, $p < 4E-07$). This and previous analysis suggest that for such an abundant class of elements as LTR is in the maize genome - and in particular in its dispensable fraction - only a subset of families have contributed to the genome expansion and the timing of insertions differs between different families.

However, this occurs mostly on intergenic fraction of the genome, as we confirmed previous studies (Kronmiller and Wise, 2009) showing that TEs found in gene islands are lower in comparison to TEs found in an “ocean” of repeats (SanMiguel et al., 1998), and that is possibly due to the fact that extensive gene disruption may potentially cause disadvantageous mutations for the plant.

4.7 SV validation based on de novo assembly

Deletions and insertions of 3 MM lines (HP301, A632, and H99) were validated on the corresponding de novo assemblies (Table 10). Validation was limited to homozygous SVs, which are reconstructed with higher confidence in de-novo assembly.

Table 10a: Validation summary of deletions

<i>Line</i>	<i>Deletions</i>	<i>Evaluated</i>	<i>Validated</i>	<i>Validated (%)</i>
HP301	23160	20257	19064	94.11%
A632	17217	14763	13548	91.77%
H99	21355	19021	16861	88.64%

Table 10b: Validation summary of insertions

<i>Line</i>	<i>Insertions</i>	<i>Evaluated</i>	<i>Validated</i>	<i>Validated (%)</i>
HP301	25307	6262	5923	94.59%
A632	18819	4543	4303	94.72%
H99	24351	6752	6332	93.78%

We were able to evaluate approximately 85% up to 89% of homozygous deletions and 24% up to 27% of homozygous insertions, validating most of them.

We were able to validate approximately 90% of the deletions (min 88.64% in H99 line and max 94.11% in HP301 line) and approximately 90 % of the insertions (min 93.78% in H99 line and max 94.72% in A632 line). This suggests that (at least for the evaluated SVs) our approaches have a low false positive rate. The investigation of false negative rate is not practical with this approach since the de-novo assembly has a very high chance of not fully reconstructing regions harboring TEs (leaving them, in the best case scenario as regions of unidentified bases between contigs or as unassembled gaps between scaffolds), as is shown by the fact that only a low proportion of insertions could be evaluated (the lack of evaluation was due to the fact that the insertion breakpoint was at the extremity of a contig).

5 CONCLUSIONS

Zea mays pan- genome composition was estimated by considering the fraction of genome shared by all the studied subjects (core genome) and the fraction that is missing in at least one of the samples (dispensable genome). The relative contribution of the dispensable and core genome to the genome of the seven lines included in the study is depicted in Figure 28.

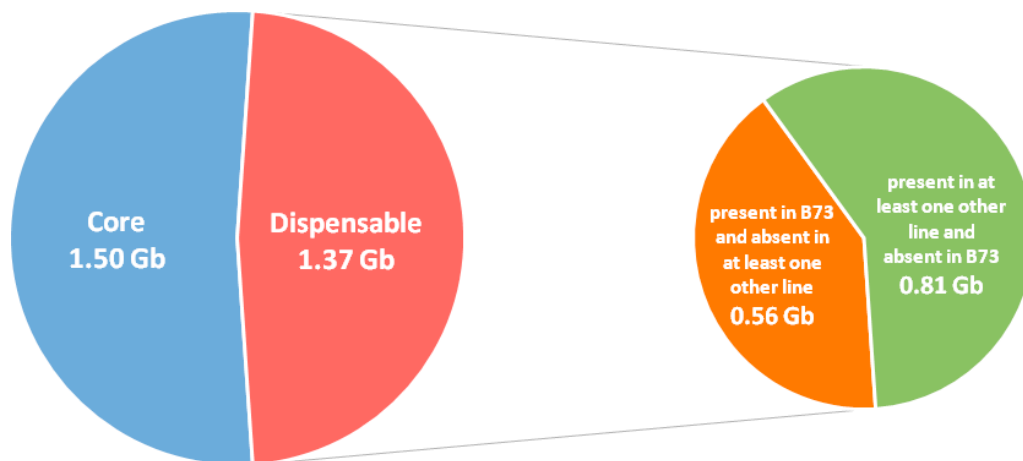


Figure 28. Pan-genome estimated size and composition (Gigabases).

Large pie: blue, core genome; red, dispensable genome. Small pie: green, insertions; orange, deletions.

About half of the pan-genome is composed by non-shared sequences. Such sequences are largely a result of a transposable element movement, and in particular to a relatively recent (3 MYA) LTR retrotransposons explosion (SanMiguel et al., 1998). We confirmed at a genome-scale level previous findings that most of LTR expansion occurred in the last half-million of years (Brunner et al., 2005). Moreover, LTR abundance in the dispensable genome was estimated to be concentrated to a small set of families, as 4 of them - equally distributed in RLC and RLG superfamilies - represent half of the total LTR fraction, which is 75%

of the B73 reference genome (Schnable et al., 2009). This is in line with previous estimations of genomic fraction of such LTR families in the maize genome, each making up 11% to 13% of the total genome sequence (Meyers et al., 2001; Kronmiller and Wise, 2008). An abundance of same families was also observed for nested elements, where defined nesting patterns exist, suggesting a sort of preference of a family to nest into another, as already observed for eukaryotic genomes (Gao et al., 2012).

As expected, most of SVs observed are intergenic. This directly correlates with the fact that most of the dispensable genome is composed by non-genic fraction. However, an analysis of genes disrupted by SVs might help to understand their phenotypic impact (Lisch et al., 2013).

The present work showed that 20% of genes are affected by a SV, while previous similar analysis report different results of 4% (Springer et al., 2009) and 60~70% (Hirsch et al., 2014; Chia et al., 2012), respectively. However, it should be considered that those results are not strictly comparable; in fact, they all used different approaches such as whole-genome sequencing, CGH array, transcriptome sequencing, and a different number of maize lines.

Springer et al. analysis is based on CGH array and can be affected by intrinsic limitations of that methodology on repeated sequences (Redon et al., 2009); this can be crucial in the maize genome, which is largely composed by repeated elements. Moreover, they performed the analysis only on 2 lines, and CGH array also lacks in resolution, while methods used in the present work can detect SV with single-base precision. Both Hirsch et al. and Chia et al. results may suggest that the analysis in the present work could be further improved by including more maize lines in the pan-genome estimation. In fact, the first performed a transcriptome analysis on 503 lines, while the latter performed a read-depth analysis on whole-genome sequencing of 103 lines.

Similarities can be found at functional level: same overrepresented GO terms are identified in both the present work and the one by Hirsch et al., in particular membrane associated terms, but also nucleotide binding, catalytic activity, and kinase associated terms.

Moreover, while considering only genes disrupted by a SV detected both here and in Springer et al. work (stringent dataset of 180 genes), it results in a subset of 43 genes; by performing an enrichment analysis, the most overrepresented GO term resulted to be “response to stress” ($p < 0.01$), which is also reported in Chia et al. analysis.

It is important to note that the observed enrichment for both “response to stress” and “membrane” related genes may be due to an enrichment for these types of genes in tandem arrays, as already noted by Rizzon et al. (2006) and observed by Swanson-Wagner et al. (2010) in another analysis performed in maize and teosinte.

A natural extension of the present work is a detailed analysis of the genes affected by structural variation, and the investigation of their function and of their possible phenotypic effects.

A large fraction of sequences was estimated to be absent in the reference while part of the dispensable genome. However evaluation of the real size of inserted sequences is difficult, since dimension of inserted sequences can only be inferred by the dimension of the reconstructed insertion (if available). The situation is further complicated by the abundance of nesting structures. To this end, a great advance may come from de novo assemblies, although the highly repetitive fraction of the maize genome can make it challenging (Treangen and Salzberg, 2011; Schatz et al., 2012). New efforts were already done in order to workaround such issues, with the aid of genotyping-by-sequencing technology coupled to machine learning (Lu et al., 2016), or PacBio single-molecule long reads (Dong et al., 2015; Wang et al., 2015). It is very recent the pre-release of the first B73

(RefGen_V4) reference genome assemblies performed using PacBio, and the genome of another elite inbred maize line (PH207) was recently assembled with the GBS anchoring pipeline (Hirsch et al., 2016).

A new algorithm to reveal insertion breakpoints was developed and used together with existing methods for structural variants discovery. Moreover, it was shown that nesting events are an important portion of the maize dispensable genome, and they were further analyzed. Nesting patterns are discovered and seen involving LTR retrotransposons above all, while a large portion of more complex patterns exists, which can be accounted to nested clusters (Kronmiller and Wise, 2009).

REFERENCES

- Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974–984 (2011).
- Aibar, S., Fontanillo, C., Droste, C. & De Las Rivas, J. Functional Gene Networks: R/Bioc package to generate and analyse gene networks derived from functional enrichment and clustering. *Bioinformatics* **31**, 1686–1688 (2015).
- Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–76 (2011).
- Andrews, S. FastQC: A quality control tool for high throughput sequence data. *Babraham Bioinforma.* 1 (2010). doi:citeulike-article-id:11583827
- Baker, M. Structural variation: the genome’s hidden architecture. *Nat. Methods* **9**, 133–7 (2012).
- Baucom, R. S. *et al.* Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genet.* **5**, (2009).
- Bennetzen, J. L. Transposable elements, gene creation and genome rearrangement in flowering plants. *Current Opinion in Genetics and Development* **15**, 621–627 (2005).
- Bennetzen, J. L. & Hake, S. *Handbook of maize: Genetics and genomics. Handbook of Maize: Genetics and Genomics* (2009). doi:10.1007/978-0-387-77863-1
- Benson, G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
- Bruggmann, R. *et al.* Uneven chromosome contraction and expansion in the maize genome. *Genome Res.* **16**, 1241–1251 (2006).
- Brunner, S., Fengler, K., Morgante, M., Tingey, S. & Rafalski, A. Evolution of DNA sequence nonhomologies among maize inbreds. *Plant. Cell.* **17**, 343–360 (2005).
- Butelli, E. *et al.* Retrotransposons Control Fruit-Specific, Cold-Dependent Accumulation of Anthocyanins in Blood Oranges. *Plant Cell* **24**, 1242–1255 (2012).
- Cao, J. *et al.* Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.* **43**, 956–963 (2011).
- Chaisson, M. J. P., Wilson, R. K. & Eichler, E. E. Genetic variation and the de novo assembly of human genomes. *Nat. Rev. Genet.* **16**, 627–640 (2015).

- Chandler, V. L. & Brendel, V. The Maize Genome Sequencing Project. *Plant Physiol.* **130**, 1594–1597 (2002).
- Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6**, 677–681 (2009).
- Chia, J. M. *et al.* Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.* **44**, 803–807 (2012).
- Del Fabbro, C., Scalabrin, S., Morgante, M. & Giorgi, F. M. An extensive evaluation of read trimming effects on illumina NGS data analysis. *PLoS One* **8**, (2013).
- Dell’Acqua, M. *et al.* Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in *Zea mays*. *Genome Biol.* **16**, 167 (2015).
- Devos, K. M., Brown, J. K. M. & Bennetzen, J. L. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res.* **12**, 1075–1079 (2002).
- Doebley, J. F., Gaut, B. S. & Smith, B. D. The Molecular Genetics of Crop Domestication. *Cell* **127**, 1309–1321 (2006).
- Dong, J. *et al.* Analysis of tandem gene copies in maize chromosomal regions reconstructed from long sequence reads. *Proc. Natl. Acad. Sci.* **113**, 7949–7956 (2016).
- Doolittle, W. F. & Sapienza, C. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**, 601–603 (1980).
- Dooner, H. K. & He, L. Maize genome structure variation: interplay between retrotransposon polymorphisms and genic recombination. *Plant Cell* **20**, 249–258 (2008).
- Du, C., Fefelova, N., Caronna, J., He, L. & Dooner, H. K. The polychromatic Helitron landscape of the maize genome. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 19916–19921 (2009).
- Eichten, S. R. *et al.* B73-Mo17 near-isogenic lines demonstrate dispersed structural variation in maize. *Plant Physiol.* **156**, 1679–90 (2011).
- Eichler, E. E. & Sankoff, D. Structural dynamics of eukaryotic chromosome evolution. *Science* **301**, 793–7 (2003).
- Eickbush, T. H. & Malik, H. S. *Origins and evolution of retrotransposons. Mobile DNA II* Edited by: Craig NL, Craigie R, Gellert M, Lambowitz AM. Washington, DC: ASM Press **93**, (2002).
- Falchi, R. *et al.* Three distinct mutational mechanisms acting on a single gene underpin the origin of yellow flesh in peach. *Plant J.* **76**, 175–187 (2013).

- Feuk, L., Carson, a R. & Scherer, S. W. Structural variation in the human genome. *Nat Rev Genet* **7**, 85–97 (2006).
- Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering transposable element diversification in de novo annotation approaches. *PLoS One* **6**, (2011).
- Fu, H. & Dooner, H. K. Intraspecific violation of genetic colinearity and its implications in maize. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 9573–9578 (2002).
- Gaut, B. S. *et al.* Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 10274–10279 (1996).
- Gierl, a & Frey, M. Eukaryotic transposable elements with short terminal inverted repeats. *Curr. Opin. Genet. Dev.* **1**, 494–7 (1991).
- Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 1513–8 (2011).
- Gore, M. A. *et al.* A First-Generation Haplotype Map of Maize. *Science (80-.)*. **326**, 1115–1117 (2009).
- Hart, S. N. *et al.* Softsearch: Integration of multiple sequence features to identify breakpoints of structural variations. *PLoS One* **8**, (2013).
- Havecker, E. R., Gao, X. & Voytas, D. F. The diversity of LTR retrotransposons. *Genome Biol.* **5**, 225 (2004).
- Hénaff, E., Zapata, L., Casacuberta, J. M. & Ossowski, S. Jitterbug: somatic and germline transposon insertion detection at single-nucleotide resolution. *BMC Genomics* **16**, 768 (2015).
- Hirsch, C. N. *et al.* Insights into the maize pan-genome and pan-transcriptome. *Plant Cell* **26**, 121–35 (2014).
- Hirsch, C. *et al.* Draft Assembly of Elite Inbred Line PH207 Provides Insights into Genomic and Transcriptome Diversity in Maize. *Plant Cell* tpc.00353.2016 (2016). doi:10.1105/TPC.16.00353
- Huang, X. & Madan, A. CAP3: A DNA sequence assembly program. *Genome Res.* **9**, 868–877 (1999).
- Jiang, N. & Wessler, S. R. Insertion preference of maize and rice miniature inverted repeat transposable elements as revealed by the analysis of nested elements. *Plant Cell* **13**, 2553–64 (2001).
- Jin, M. *et al.* Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation. *Scientific Reports* **6**, 1–12 (2016). doi:10.1038/srep18936

- Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
- Keane, T. M., Wong, K. & Adams, D. J. RetroSeq: Transposable element discovery from Illumina paired-end sequencing data. *Bioinformatics* 1–2 (2012). doi:10.1093/bioinformatics/bts697
- Kersey, P. J. *et al.* Ensembl Genomes: An integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Res.* **40**, (2012).
- Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide-sequences. *J Mol Evol* **16**, 111–120 (1980).
- Kobayashi, S., Goto-Yamamoto, N. & Hirochika, H. Retrotransposon-induced mutations in grape skin color. *Science* **304**, 982 (2004).
- Kronmiller, B. A., Wise, R. P. & Walker, J. M. in *Plant transposable elements: Methods and protocols* **1057**, 305–319 (2013).
- Kronmiller, B. A. & Wise, R. P. Computational finishing of large sequence contigs reveals interspersed nested repeats and gene islands in the *rfl1*-associated region of maize. *Plant Physiol.* **151**, 483–495 (2009).
- Kronmiller, B. A. & Wise, R. P. TENest: automated chronological annotation and visualization of nested plant transposable elements. *Plant Physiol.* **146**, 45–59 (2008).
- Lai, J. *et al.* Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat. Genet.* **42**, 1027–1030 (2010).
- Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).
- Levin, H. L. & Moran, J. V. Dynamic interactions between transposable elements and their hosts. *Nat. Rev. Genet.* **12**, 615–627 (2011).
- Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997v2 [q-bio.GN]* (2013).
- Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
- Li, Y. & Dooner, H. K. Excision of Helitron transposons in maize. *Genetics* **182**, 399–402 (2009).
- Lisch, D. How important are transposons for plant evolution? *Nat Rev Genet* **14**, 49–61 (2012).

- Liu, S. *et al.* Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. *PLoS Genet.* **5**, (2009).
- Lu, F. *et al.* High-resolution genetic mapping of maize pan-genome sequence anchors. *Nat. Comm.* **6**, 6914 (2015).
- Ma, J. & Bennetzen, J. L. Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 12404–10 (2004).
- Magris, G. Characterisation of the pan-genome of *Vitis vinifera* using Next Generation Sequencing. (Doctoral Thesis, Università degli Studi di Udine, 2016).
- Maron, L. G. *et al.* Aluminum tolerance in maize is associated with higher MATE1 gene copy number. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 5241–5246 (2013).
- Marroni, F., Pinosio, S. & Morgante, M. Structural variation and genome complexity: Is dispensable really dispensable? *Current Opinion in Plant Biology* **18**, 31–36 (2014).
- Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* **17**, 10–12 (2011).
- McClintock, B. Chromosome organization and genic expression. *Cold Spring Harb. Symp. Quant. Biol.* **16**, 13–47 (1951).
- McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297–1303 (2010).
- Medvedev, P., Stanciu, M. & Brudno, M. Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods* **6**, S13–S20 (2009).
- Messing, J. & Dooner, H. K. Organization and variability of the maize genome. *Current Opinion in Plant Biology* **9**, 157–163 (2006).
- Meyers, B. C., Tingey, S. V. & Morgante, M. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Res.* **11**, 1660–1676 (2001).
- Morgante, M. *et al.* Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat. Genet.* **37**, 997–1002 (2005).
- Morgante, M., De Paoli, E. & Radovic, S. Transposable elements and the plant pan-genomes. *Current Opinion in Plant Biology* **10**, 149–155 (2007).
- Nourimand, M. & Todd, C. D. Allantoin Increases Cadmium Tolerance in *Arabidopsis* via Activation of Antioxidant Mechanisms. *Plant Cell Physiol* (2016).

- Paterson, a H., Bowers, J. E. & Chapman, B. a. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 9903–8 (2004).
- Piegu, B. *et al.* Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.* **16**, 1262–1269 (2006).
- Pinosio, S. Building catalogues of genetic variation in Poplar. (Doctoral Thesis, Università degli Studi di Udine, 2012).
- Pinosio, S. *et al.* Characterization of the Poplar Pan-Genome by Genome-Wide Identification of Structural Variation. *Mol. Biol. Evol.* **33**, 2706–2719 (2016).
- Pritham, E. J. & Thomas, J. Helitrons, the Eukaryotic Rolling-circle Transposable Elements. *Microbiol. Spectr.* **3**, 1–32 (2015).
- Rausch, T. *et al.* DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, (2012).
- Redon, R. & Carter N. P. Comparative Genomic Hybridization: microarray design and data interpretation. *Methods Mol. Biol.* **529**, 37–49 (2009).
- Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
- Rizzon C. *et al.* Striking Similarities in the Genomic Distribution of Tandemly Arrayed Genes in Arabidopsis and Rice. *PLoS Comput. Biol.* **9** (2006).
- Sankoff, D. & Zheng, C. Fractionation, rearrangement and subgenome dominance. *Bioinformatics* **28**, (2012).
- SanMiguel, P., Gaut, B. S., Tikhonov, a, Nakajima, Y. & Bennetzen, J. L. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20**, 43–45 (1998).
- SanMiguel, P. *et al.* Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274**, 765–768 (1996).
- Schatz, M. C., Witkowski, J. & McCombie, W. R. Current challenges in de novo plant genome sequencing and assembly. *Genome Biol.* **13**, 243 (2012).
- Schnable, P., Ware, D., Fulton, R. & Stein, J. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009).
- Selinger, D. A. & Chandler, V. L. B-Bolivia, an allele of the maize b1 gene with variable expression, contains a high copy retrotransposon-related sequence immediately upstream. *Plant Physiol.* **125**, 1363–1379 (2001).
- Shamos, M. I. & Hoey, D. *Geometric intersection problems. 17th Annual Symposium on Foundations of Computer Science (sfcs 1976)* (1976). doi:10.1109/SFCS.1976.16

- Sindi, S., Helman, E., Bashir, A. & Raphael, B. J. A geometric approach for classification and comparison of structural variants. *Bioinformatics* **25**, i222–30 (2009).
- Smit, A., Hubley, R. & Green, P. RepeatMasker Open-3.0. *RepeatMasker Open-3.0* www.repeatmasker.org (1996).
- Springer, N. M. *et al.* Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet.* **5**, (2009).
- Studer, A., Zhao, Q., Ross-Ibarra, J. & Doebley, J. Identification of a functional transposon insertion in the maize domestication gene *tb1*. *Nat. Genet.* **43**, 1160–1163 (2011).
- Swanson-Wagner, R. A. *et al.* Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res.* **20**, 1689–1699 (2010).
- Swigoňová, Z. *et al.* Close split of sorghum and maize genome progenitors. *Genome Res.* **14**, 1916–1923 (2004).
- Tello-Ruiz, M. K. *et al.* in *Methods in Molecular Biology* **1374**, 141–163 (2016).
- Tenaillon, M. I., Hollister, J. D. & Gaut, B. S. A triptych of the evolution of plant transposable elements. *Trends in Plant Science* **15**, 471–478 (2010).
- Tenaillon, M. I., Hufford, M. B., Gaut, B. S. & Ross-Ibarra, J. Genome size and transposable element content as determined by high-throughput sequencing in maize and *Zea luxurians*. *Genome Biol. Evol.* **3**, 219–229 (2011).
- Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial ‘pan-genome’. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 13950–5 (2005).
- Thomas, C. A. The genetic organization of chromosomes. *Annu. Rev. Genet.* **5**, 237–256 (1971).
- Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* **46**, 36–46 (2012).
- Wang, B. *et al.* Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Comm.* **7**, 11708 (2016).
- Wang, J. *et al.* CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat. Methods* **8**, 652–4 (2011).
- Wang, Q. & Dooner, H. K. Remarkable variation in maize genome structure inferred from haplotype diversity at the *bz* locus. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 17644–9 (2006).

- Wang, X., Weigel, D. & Smith, L. M. Transposon Variants and Their Effects on Gene Expression in Arabidopsis. *PLoS Genet.* **9**, (2013).
- Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).
- Xiong, W., He, L., Lai, J., Dooner, H. K. & Du, C. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proc. Natl. Acad. Sci.* **111**, 10263–10268 (2014).
- Xu, Z. & Wang, H. LTR-FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, (2007).
- Yang, L. & Bennetzen, J. L. Structure-based discovery and description of plant and animal Helitrons. *Proc. Natl. Acad. Sci.* **106**, 12832–12837 (2009).
- Yang, L. & Bennetzen, J. L. Distribution, diversity, evolution, and survival of Helitrons in the maize genome. *Proc. Natl. Acad. Sci.* **106**, 19922–19927 (2009).
- Zhao, D. & Jiang, N. Nested insertions and accumulation of indels are negatively correlated with abundance of Mutator-like transposable elements in maize and rice. *PLoS One* **9**, (2014).

Additional notes

Figures 1-8 (present in the Introduction chapter) were taken from published journals with permission of relative editors, and/or in any case in the respect of the Italian copyright law Art. 70 paragraph 1 (ii) which grants the reuse of low-quality figures for scientific non-commercial purposes, so called “fair use”.

Custom scripts and unpublished tools are available on request (ezapparoli@appliedgenomics.org).

ACKNOWLEDGMENTS

I would like to thank Prof. Michele Morgante who coordinated the work and offered me the opportunity to carry out this PhD thesis with the full support of the Applied Genomic Institute.

I would also thank my co-supervisor Fabio Marroni, for his valuable advice and lessons concerning genomic and statistical matters, and for the scientific support in general.

In particular, for Walle development, I would like to thank: Cristian Del Fabbro - who gave me the idea for the sweep-line parsing implementation - for his more than valuable algorithmic thinking; Davide Scaglione for his precious advice and experience in genomic application of computer programming; Vittorio Zamboni who taught me Python and helped me to write the very first version of the Walle codebase.

Moreover, I would like to thank in particular Gabriele Magris, Michele Vidotto and Sara Pinosio, for their essential help in performing a large part of analysis, making available to me ad hoc algorithms and pipelines developed by them.

Furthermore, I thank my present and former colleagues at the Applied Genomic Institute and the University of Udine, for the inspiration and scientific support in the last 3 years: Aldo Tocci, Alessandro Gervaso, Alice Fornasiero, Andrea Zuccolo, Eleonora Paparelli, Emanuele De Paoli, Gabriele Di Gaspero, Mara Miculan, Mirko Celii, Rachel Schwope, Simone Scalabrin, Vera Vendramin and all the lab people.

The present work has been supported by the European Commission's European Research Council, within the Seventh Framework Programme for Research (Grant number, 294780).